SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL
3ºONTOBRAS

Construindo ontologias de domínio: o (re)conhecimento da intensificação agropecuária no Brasil.

Ivo Pierozzi Jr. (Embrapa Informática Agropecuária) ivo@cnptia.embrapa.br

Leandro HenriqueMendonça de Oliveira (Embrapa Informática Agropecuária) leandro@cnptia.embrapa.br

Kleber Xavier Sampaio de Souza (Embrapa Informática Agropecuária) kleber@cnptia.embrapa.br

Resumo: Em seus projetos de PD&I, a Embrapa necessita integrar informações oriundas dos mais variados domínios de conhecimento, visando soluções para os inúmeros problemas que afetam a atividade agrícola nacional. No caso do entendimento da intensificação agropecuária, observada em algumas regiões produtoras de commodities agrícolas, propõe-se o desenvolvimento de estudos terminológicos como suporte ao processo de organização, disseminação e apropriação do conhecimento, tanto pela comunidade científica como pelos usuários das tecnologias, serviços e produtos da Embrapa. Nesse contexto, o presente artigo relata o desenvolvimento de um trabalho de construção de um mapa conceitual focado no tema da intensificação agropecuária. Esse trabalho tem como suporte o uso de ferramentas computacionais e tecnologias de informação para tratamento semi-automático dos termos e construção de taxonomia e ontologia específicas.
Palavras-chave: Embrapa; Mapa Conceitual; Extração Automática de Termos; Mineração de Textos; Gestão do Conhecimento.

1. Introdução

No mosaico territorial da agropecuária brasileira, a dinâmica do uso e cobertura das terras envolve transições que ainda ocorrem descompassadas com as iniciativas de planejamento territorial. Na miríade de situações atuais observam-se processos de expansão agrícola simultaneamente aos de intensificação agropecuária. Justamente por ser complexa, a discussão integrada sobre territorialidade, sustentabilidade e competitividade das atividades agropecuárias no Brasil necessita de abordagens parciais, mas que não se percam do contexto geral e sejam passíveis de integrar conclusões e resultados de outras abordagens, viabilizando planejamento e implantação de políticas públicas eficientes e eficazes.

Nesse contexto, a Embrapa Informática Agropecuária, centro temático de tecnologia da informação da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), sediado em Campinas, SP, desenvolve atualmente o projeto INTAGRO (*Intensificação Agropecuária em Pólos de Produção de Soja e Cana-de-Açúcar: Territorialidade, Sustentabilidade e Competitividade*), que propõe a integração de várias abordagens visando a busca de respostas na visão integrada de vários planos. No plano agroambiental são diagnosticados, georreferenciados e cartografados os processos de intensificação agropecuária, em áreas de produção das principais *commodities* agrícolas brasileiras, notadamente soja e cana-de-açúcar,

relacionando tais processos às características e à conservação dos solos, à produtividade das culturas e à ocorrência de queimadas, para um melhor entendimento das influências sofridas por aqueles processos ou das consequências originadas por eles. Já o plano socioeconômico, visa o melhor entendimento de determinantes e tendências que modulam e direcionam a intensificação agropecuária naqueles setores de produção agrícola.

Mas é no plano do avanço, organização e disseminação do conhecimento, que o projeto mais inova, ao propor a realização de estudos e análises integradas dos dados e construção de cenários futuros. Para tanto, estão sendo desenvolvidas atividades de conceitualização e categorização terminológica para subsidiar o processo de integração e apropriação dos conhecimentos gerados. Esses estudos focam a construção de taxonomias (e, posteriormente, ontologias) para, principalmente, criar um arranjo dos conceitos desse domínio, servindo para organizar os campos nocionais do mesmo, de modo que se possa compreender suas hierarquias e interrelações conceituais básicas. Além disso, busca-se assegurar que os resultados do projeto possam ser relacionados consistentemente a diferentes iniciativas dentro desse domínio de conhecimento e, assim, possam contribuir para que dados e informações sejam adequadamente relacionados e recuperados, servindo eficazmente para a formulação de estratégias de desenvolvimento e de políticas públicas que visem melhorias de sustentabilidade e competitividade do setor agropecuário brasileiro.

2. Contextualização do trabalho e referencial teórico

Esta seção não representa uma revisão exaustiva de literatura e sim uma apresentação resumida de estudos e trabalhos que abrangem o domínio de conhecimento aqui considerado, com intuito de: (1) contextualizar a problemática do entendimento e uso do termo "intensificação agropecuária"; (2) contextualizar como são trabalhadas as terminologias, taxonomias e ontologias de domínio e, consequentemente, como esse trabalho pode ser usado como prática de gestão do conhecimento e (3) contextualizar o uso de tecnologias da informação como ferramentas de suporte e facilitação do trabalho terminológico.

2.1 O termo "intensificação agropecuária"

De maneira genérica, pode-se entender a intensificação agropecuária como qualquer prática agropecuária que aumenta a produtividade por unidade de área. Desse modo, intensificação agropecuária pode representar desde redução no período de pousio e multissafras até irrigação, fertilização, uso de animais de carga, maquinaria, variedades de vegetais ou animais modificados geneticamente, defensivos químicos, etc.

O termo "intensificação agropecuária" é utilizado neste trabalho, originado da tradução literal do inglês *agricultural intensification* (BOSERUP, 1965). Terminologicamente falando, embora a autora tenha desenvolvido claramente o conceito ao qual o termo se refere o mesmo não foi "formalmente definido" na citada obra. Na concepção dessa economista dinamarquesa, a "intensificação agropecuária" é um processo agrossocioeconômico complexo e se contrapõe à visão malthusiana de que o aumento populacional leva à escassez de alimentos. Para Boserup, a pressão do aumento populacional, num dado território, induz a intensificação dos processos agropecuários e estes, por sua vez, acabam por suprir as necessidades da população crescente.

A partir de Boserup, o que se observa na literatura é que o termo passou a ser utilizado de forma imprecisa e por vezes ambígua, afastando-se do conceito agrossocioeconômico inicialmente proposto. Assim, muitas vezes, a literatura utiliza "intensificação agropecuária" para se referir a outros processos como modernização ou tecnicização ou até mesmo expansão agrícola sobre áreas ainda não utilizadas, com cobertura vegetal nativa. Vale ressaltar duas considerações em relação à necessidade de um melhor entendimento do termo "intensificação agropecuária": (1) a questão relacionada aos problemas advindos da sua tradução da língua inglesa para a portuguesa; (2) a recuperação do sentido e da lógica conceituais, inicialmente propostos por Boserup, respeitando toda a complexidade interdisciplinar envolvida.

É com base nesse contexto que se justifica a construção de ontologia para organização do conhecimento deste domínio específico, já que elas correspondem, portanto, a uma representação de um domínio a partir de seus conceitos abstratos e a forma como eles se relacionam. Ou seja, um modelo consensual de um mundo particular, reconhecido da mesma forma pelas pessoas desse mundo.

2.2 Ontologias e tratamento terminológico

Uma ontologia ou estrutura conceitual pode ser entendida como "uma especificação de uma conceitualização" (GRUBER, 1993), ou seja, um modelo comum ou estrutura conceitual sistematizada e de consenso que permite não só armazenar, mas também buscar e recuperar a informação sobre um determinado domínio do conhecimento. Mais especificamente, uma ontologia define os termos e as relações básicas para a compreensão de uma área do conhecimento.

Nesse contexto, a ontologia na pesquisa terminológica é fundamental para: (1) possibilitar um mapeamento mais sistemático de um campo de especialidade; (2) circunscrever a pesquisa, já

que todas as ramificações da área-objeto, com seus campos, são previamente mapeadas; (3) delimitar o conjunto terminológico; (4) determinar a pertinência dos termos, pois separando cada grupo de termos pertencentes a um determinado campo, pode-se apontar quais termos são relevantes para o trabalho e quais não são; (5) prever os grupos de termos pertencentes à área-objeto, como também os que fazem parte de matérias conexas; (6) definir as unidades terminológicas de maneira sistemática e, finalmente, (7) controlar a rede de remissivas (ALMEIDA, 2000).

A construção de ontologias de domínio, no entanto, é tarefa pouco trivial, já que os processos que a compõem (a saber: extração dos candidatos a termos e posterior alocação dos mesmos em uma estrutura conceitual), são tradicionalmente manuais e, portanto, demorados (ALMEIDA *et. al.*, 2006). Dessa forma, é natural que se busquem métodos para automatizar ou semi-automatizar um ou mais dos processos de construção de ontologias. Uma das maneiras de semi-automatizar esse processo é utilizar um *corpus* textual. Para isso, toma-se por base princípios teórico-metodológicos advindos de duas áreas de pesquisa: da Lingüística, mais especificamente da Terminologia (ALMEIDA, 2000; CABRÉ, 1993; 1999), e do Processamento Automático das Línguas Naturais (PLN), com ênfase no uso de ferramentas para (1) a montagem de *corpus* e extração automática de termos e (2) definição da própria ontologia (ALMEIDA *et. al.*, 2006). Genericamente, pode-se chamar de *corpus* qualquer coleção de textos selecionados para caracterizar uma língua ou variedade de língua, de uma área específica ou não do conhecimento humano (SARDINHA, 2004).

2.3 Metodologias e ferramentas para tratamento terminológico

A metodologia semi-automática divide a tarefa de construção de ontologias em duas partes dependentes entre si, sendo uma humana (manual) e outra automática, utilizando-se de ferramentas computacionais específicas. Se por um lado, essa metodologia compartilha com a manual o trabalho colaborativo com os especialistas da área cuja ontologia está sendo elaborada, por outro lado, difere da manual quanto ao tempo despendido, principalmente nas tarefas de compilação do *corpus*, extração dos candidatos a termos, e quanto ao uso de critérios estatísticos para a extração desses candidatos. As etapas constitutivas da metodologia semi-automática e as estratégias/ferramentas adotadas em cada uma delas são detalhadas por Di Fellipo *et. al.* (2009) e foram utilizadas neste trabalho exatamente como descritas por esses autores. A junção das duas etapas (manual e semi-automática) permite a construção de uma ontologia representando uma visão estruturada do domínio do conhecimento e permite o compartilhamento de informação com maior eficácia.

3. Material e métodos

No trabalho aqui relatado, a delimitação da área de conhecimento partiu da escolha em se trabalhar com o termo em inglês "*agricultural intensification*". Essa escolha se deveu à fraca recuperação de literatura disponível em português a partir da utilização do bigrama "*intensificação agropecuária*", como elemento de busca em bases de dados bibliográficos ou até mesmo na Internet e que resultou em expressiva quantidade de referências classificadas como literatura cinzenta. Além disso, considerando-se, ainda, o unigrama "*intensificação*" como palavra de pesquisa, constatou-se uma enorme variação de sentidos, indicando a utilização imprecisa e generalizada do termo no domínio do conhecimento pretendido. Dessa forma, o *corpus* foi construído utilizando-se o bigrama "*agricultural intensification*" como elemento de busca de referências bibliográficas. Utilizou-se para essa busca o *software* EndNotes X1, optando-se pelo acesso remoto e direto à base bibliográfica *Web of Science* (*ISI*). Esse processo de busca resultou na recuperação de referências bibliográficas em língua inglesa, constando tanto artigos integrais como resumos. A maioria dos artigos integrais foram recuperados no formato PDF os quais, juntamente com o conteúdo dos resumos, foram convertidos em arquivos no formato de texto puro (extensão .TXT, sem formatação), para a extração automática dos termos. O resultado deste processo obteve 1280 referências bibliográficas, abrangendo o período 1964-2009; desse total, foram efetivamente usadas 393 delas na forma de texto completo e 283 na forma de resumos, formando um *corpus* com 2.570.923 palavras.

Depois da compilação do *corpus*, procedeu-se a extração automática de candidatos a termos. A ferramenta utilizada foi o NSP (Ngrams Statistic Package) (PEDERSEN *et. al.*, 2003), um conjunto de programas desenvolvido para identificar e extrair n-grams (uma sequência contínua de palavras) do *corpus* com os seguintes parâmetros: (1) tamanho do n-gram; (2) stoplist**:** correspondendo a uma lista de palavras que devem ser ignoradas durante o processamento, incluindo palavras funcionais com alta freqüência que não apresentam valor terminológico como: preposições, artigos, conjunções e advérbios e palavras de marcação estrutural dos textos técnicos/científicos como "Introdução", "Referências", "Bibliografia", etc; (3) ponto de corte: limiar inferior para o qual termos com uma frequência absoluta menor serão desprezados. Neste trabalho o corte foi definido em 24; (4) regra de formação de palavras: que permite a definição e especificação de qual padrão de palavras deve ser selecionado durante sua execução. Para este trabalho, a regra de formação foi definida para aceitar palavras maiúsculas e minúsculas do idioma inglês.

4. Resultados preliminares

O primeiro resultado preliminar produzido por este trabalho é o próprio *corpus* textual, já que sua compilação representa um recurso linguístico rico e autêntico que pode ser usado em outras pesquisas, tanto para confecção de ontologias como naquelas de mineração de texto e processamento de língua natural.

O segundo e principal resultado desta pesquisa é a taxonomia criada a partir dos *n-grams* extraídos. Nesta primeira versão, foram identificados e organizados hierarquicamente 638 nós conceituais distribuídos em 7 conceitos primários (primeiro nível da hierarquia). O processo de extração dos *n-grams* produziu 6630 unigramas, 1687 bigramas e 2781 trigramas. Nesta fase do trabalho, a elaboração da taxonomia considerou somente os bigramas, uma vez que a análise dos *n-grams* extraídos concluiu que os unigramas denotavam um universo de conceitos demasiadamente genéricos e os trigramas repetiam conceitos já cobertos pelos bigramas. Entretanto, vale salientar que tal decisão não os exclui e os mesmos serão retomados, posteriormente, para um ajuste mais refinado da taxonomia. Prevê-se também que na fase de validação final da taxonomia, muitos outros termos tenham que ainda ser excluídos e outros, não extraídos do *corpus*, incorporados. A Figura 1 mostra parte da primeira versão da taxonomia construída a partir lista de alguns termos extraídos do corpus (1a) e os mesmos termos alocados na estrutura na visualizações *folder-tree* (1b) e hiperbólica (1c).

criado uma ferramenta de visualização computacional dinâmica que mostra os termos, seus respectivos equivalentes e a relação entre eles em uma rede de nós e arestas que pode ser movida dinamicamente por toda a base. A Figura 2 apresenta uma tela dessa visualização destacando o termo ***"land cover"*** e seus relacionamentos. Observe que os termos são apresentados pelos retângulos e as arestas as relações entre eles. Outra vantagem dessa visualização é a possibilidade de utilizá-la para representar também a taxonomia criada.

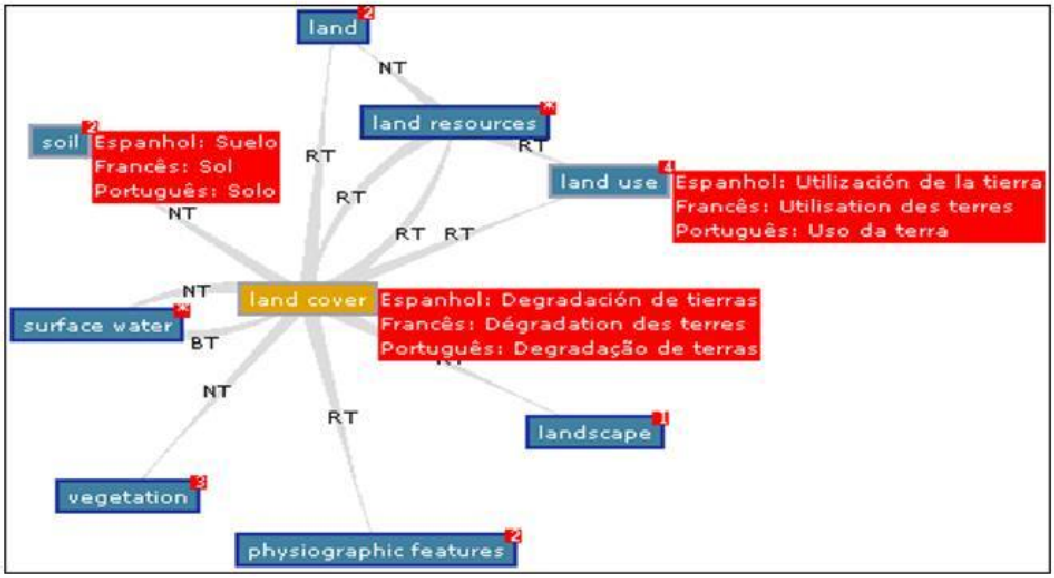

Figura 2 – Visualização da comparação do termo ***"land cover"*** com o Agrovoc Thesaurus.

Como quarto e último resultado, destacamos o desenvolvimento da ferramenta *WEAT (Extrator Web Automático de Termos)*, um programa que pode ser utilizado gratuitamente via internet para a extração automática de termos. Nela o usuário envia seu *corpus* (operação de *upload)* e por meio da seleção de vários parâmetros define a maneira que deseja extrair seus candidatos a termos. Além disso, o WEAT oferece recursos computacionais para análise de *corpus* como, por exemplo, o *concordanciador,* uma ferramenta que consulta e apresenta o contexto de todas as ocorrências de um determinado termo no *corpus*, e os gráficos dinâmicos que mostram a distribuição das palavras ao longo do *corpus*. Nesse contexto, a vantagem primordial do WEAT é disponibilização gratuita de ferramentas e recursos destinados à extração de termos ao usuário final, democratizando este acesso e não restringindo tais funcionalidades somente aos profissionais da informática. A Figura 3 apresenta duas telas sobrepostas de algumas possibilidades de uso do WEAT.

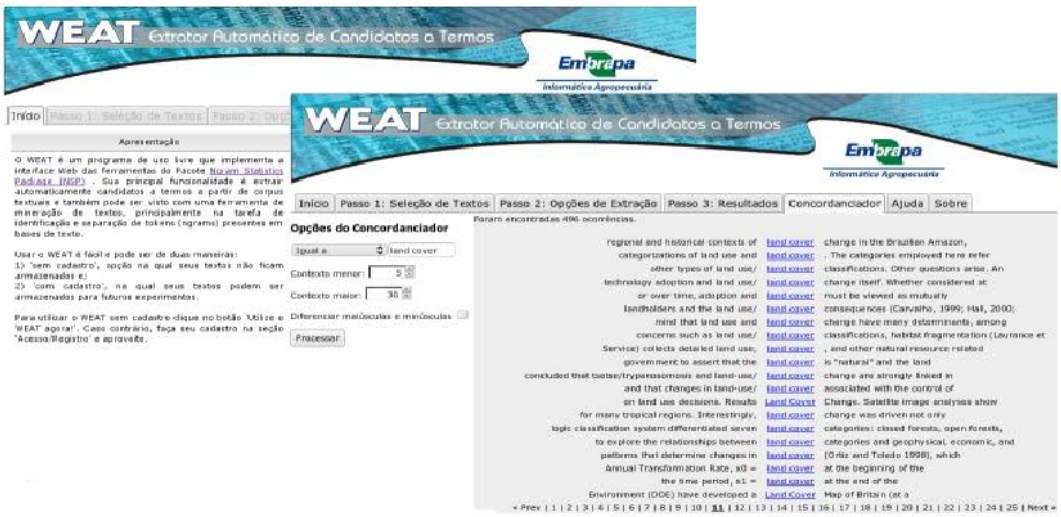

Figura 3 – Telas de alguma funcionalidades do WEAT

5. Trabalhos futuros e considerações finais

A principal iniciativa de continuidade deste trabalho é o refinamento da taxonomia desenvolvida. Pretende-se aprimorar os parâmetros da extração de termos para subsidiar a evolução da taxonomia. Outras ações são o incremento da comparação dos termos extraídos com o Agrovoc, visando descobrir o grau de cobertura dos termos extraídos em relação a este Thesaurus, e a implementação de novas versões do WEAT, com disponibilização de medidas estatísticas de associação para facilitar o julgamento, análise e seleção dos termos extraídos automaticamente. Pretende-se, ainda, exercitar e divulgar a utilidade de se utilizar procedimentos e ferramentas de gestão do conhecimento como suporte a abordagens mais complexas de PD&I. As questões relacionadas à organização e sistematização do conhecimento, beneficiando-se da elaboração de taxonomias e ontologias de domínio, podem ser um diferencial facilitador no desenvolvimento de abordagens mais complexas, como as pretendidas pelo projeto INTAGRO, entre elas as análises integradas de dados e informações e a construção de cenários influenciados por uma gama enorme de variáreis ambientais, agronômicas e socioeconômicas. Essa facilitação se revela quando se constata a complexidade que emerge de tais iniciativas, pois é necessário o entendimento e a integração de vários domínios de conhecimento para sua execução e refinamento. Assim, o exercício terminológico, aqui relatado, pretende oferecer aos próprios pesquisadores do projeto INTAGRO e, posteriormente, aos usuários do conhecimento gerado pela Embrapa: (1) uma proposta de organização do conhecimento sobre o processo de intensificação agropecuária; (2) uma representação clara das relações entre os conceitos desse domínio de conhecimento; (3) a revelação de conceitos ainda inexistentes ou conceitos redundantes (sinonímia) no

contexto desse domínio de conhecimento e (4) uma proposta de normalização da terminologia com equivalências ou não dos termos pertinentes a esse domínio de conhecimento. Algumas indicações qualitativas já podem ser ressaltadas pela análise dos termos extraídos. Por exemplo, constata-se forte presença de termos que denotam conceitos referentes aos domínios de conhecimentos com os quais o processo de "intensificação agropecuária" se relaciona, indicando que daí uma terminologia consistente possa se originar. Por exemplo, observam-se termos que descrevem ou definem conceitos relacionados a dinâmicas demográficas e do uso e cobertura das terras; territorialidade; impacto sobre a biodiversidade; sistemas de manejo agrícola e produtividade agropecuária, entre outros, resgatando o conceito original do trabalho de Boserup (1965). Com o aprimoramento dessa taxonomia e a inclusão de abordagens quantitativas adequadas para a análise terminológica, será possível evidenciar se novos termos foram criados, se alguns desapareceram ou se é necessária a inclusão de outros ausentes.

Referências

ALMEIDA, Gladis Maria Barcellos. **Teoria comunicativa da terminologia: uma aplicação**. 2000. vol. I, 290; vol. II, 86 f. Faculdade de Ciências e Letras, Campus de Araraquara, Universidade Estadual Paulista, Araraquara, Doutorado em Linguística e Língua Portuguesa.

ALMEIDA, Gladis Maria Barcellos; OLIVEIRA, Leandro Henrique Mendonça; ALUÍSIO, Sandra Maria. **A terminologia na era da informática**. Ciencia e Cultura [online], v. 58, n. 2, p. 42-45, abr./jun. 2006. Disponível em: <http://cienciaecultura.bvs.br/scielo.php?pid=0009-6725&script=sci_serial>.

Boserup, Ester. **The Conditions of Agricultural Growth: The Economics of Agrarian Change under Population Pressure**. London: Allen & Unwin, 1965.

CABRÉ, Maria Tereza. **La terminología: representacíon y comunicacíon - elementos para una teoria de base comunicativa y outros artículos**. Barcelona: Institut Universitari de Linguistica Aplicada, 1999.

CABRÉ, Maria Tereza. **La terminología: teoría, metodología, aplicaciones**. Antártida/Empúries, Barcelona, 1993.

SARDINHA, Tony Beber. **Linguística de Corpus**. Editora Manole, Barueri-SP, 2004.

OLIVEIRA, Leandro Henrique Mendonça. **e-Termos: Um ambiente colaborativo web de gestão terminológica**. Instituto de Ciências Matemáticas e de Computação (ICMC). Universidade de São Paulo (USP). Tese de Doutorado. São Carlos-SP. Setembro de 2009. 331p.

GRUBER, Thomas Robert. **A translation approach to portable ontologies**. Knowledge Acquisition, v. 5, n. 2, 1993.