AGROVOC as Knowledge Organization Model applied to Brazilian Agricultural Intensification Processes

Ivo Pierozzi Jr.[1], Leandro Henrique Mendonça de Oliveira[1], Gladis Maria de Barcellos Almeida[2], Caterina Caracciolo[3] and Gudrun Johannsen[3]

[1]*Embrapa Informática Agropecuária, Campinas, SP, Brazil*
{ivo.pierozzi, leandro.oliveira}@embrapa.br
[2]*Universidade Federal de São Carlos, São Carlos, SP, Brazil*
{gladis.mba@gmail.com}
[3]*Office of Knowledge Exchange, Research and Extension/FAO, Rome, Italy*
{caterina.caracciolo, gudrun.johannsen@fao.org}

Abstract. This paper shows the experience of using AGROVOC thesaurus as a basis and model of information and knowledge organization and domain representation applied to soybean and sugar cane agricultural intensification in Brazil. A textual corpus of about 2,5 million of words was compiled and candidate terms extracted automatically for creating an initial hierarchical conceptual map, then compared to AGROVOC terms. The results show the importance of AGROVOC as a resource to organize and represent agricultural information and knowledge, as a reference for creating new terminological products on agriculture, and open new possibilities for enrichment of AGROVOC´s Brazilian Portuguese terminologies.

Keywords: knowledge organization and representation, KOS, agricultural terminologies, agricultural intensification, Brazilian Portuguese idiomatic variations

1. Introduction

Generically, the increase of agricultural production can happen following one of two ways not necessarily mutually excludible: either territorially expanding farming surfaces or intensifying agricultural activities in the same given area through higher inputs of capital, labor and technology. The second way has been linked to the concept of agricultural intensification (AI) [1-2], which has been identified as a major process contributing to the performance of Brazilian agricultural sector in the last three decades, when it was observed repeated records of production and productivity of agricultural commodities such as grains or raw material for biofuels [3].

Opposing to this favorable scenario, the global concerns with preservation and conservation of areas still covered with natural vegetation and food security set increasingly international pressures and put Brazil on the outskirts of conducting their conventional patterns of land use and land cover, requiring effective, efficacious and sustainable solutions collectively negotiated, built on solid technical bases and consensual scientific conceptualizations.

The mission of Brazilian Agricultural Research Corporation (Embrapa) includes knowledge conversion into effective solutions to society. To achieve this goal, Embrapa has adopted collaborative networking arrangements to organize and develop its research, development and innovation (RD&I) [4-6]. But when trying to work collaboratively members belonging to these networks have difficulties for creating and sharing knowledge due to conflicts caused by barriers in the information exchange such as: geographic dispersion of people; various media formats in which information is produced and distributed; multi-disciplinary nature of knowledge, involving

teams of professionals from different specialties; linguistic differences between experts and even differences between schools of thought.

These difficulties are directly related to constituents of the cognition, meaning and communication human processes. Moreover, RD&I networks with modern patterns of functioning are highly dependent on computing resources; so such difficulties are projected beyond the natural languages reaching artificial languages used by computers. In such situation the use of formal and standardized models of knowledge organization improves both interpersonal communication and semantic interoperability of information systems. Scientific discourse and application of scientific results is then strengthened. Such models are called knowledge organization systems (KOS) [7-8]. This article reports on the experience of using AGROVOC thesaurus (http://aims.fao.org/standards/agrovoc/about) as a model of information and knowledge organization and representation applied to the context of AI in soybean and sugar cane in Brazil.

2. Agricultural Intensification: Why KOS May Help

Embrapa Agricultural Informatics, a thematic center of applied information technology to agriculture has developed the project entitled “Agricultural Intensification in Soybean and Sugar Cane Productive Areas: Territoriality, Sustainability and Competitiveness”. This project proposed different scientific approaches to study this matter by integrating various knowledge domains. Regarding knowledge organization and dissemination aspects, the project has proposed terminological categorization and conceptualization activities to support the integration and appropriation of generated knowledge. Such activities included the construction of models to create an arrangement of concepts in that domain, serving to organize its notional fields, allowing understand its basic conceptual hierarchies and interrelationships [9].

As mentioned above, one can understand “agricultural intensification” as any practice that increases agricultural production in given area unit at some cost in labor or capital inputs. Thus, AI may represent a reduction of fallow period and consequent multi harvests; intensified use of machinery, chemical pesticides, irrigation, fertilization; use of draft animals; genetically modified plant or animal varieties and so on. The term was presented by Ester Boserup in [1], however, the concept of agricultural intensification was not "formally defined" in that work. Boserup´s original concept of AI refers to a social and economic complex process. After Boserup, this term is usually found in literature used in an imprecise and sometimes ambiguous away. Often, it is used to refer to other concepts such as “intensive agriculture”, “modernization” or “technification” or even “agricultural expansion” process consisting in major agricultural productions by transformation of native vegetation into farmlands, i.e., something quite different from Boserup´s AI original conceptualization.

Based on the above argumentation, the need for a better understanding of what AI is emerges clearly: we need to conciliate the original conceptualization proposed by Boserup, with the complexity involved by its intrinsic interdisciplinarity. This backdrop justifies the application of knowledge organization systems for organizing and representing this agricultural knowledge domain.

3. A KOS Based on AGROVOC

For constructing a terminology about AI we refer to the methodology exposed in [9-10]. As a first step, we constructed an English textual corpus, by using the bigram "agricultural intensification" as keyword for searching bibliography on this subject. This choice was due to the weak recovery of available literature in Portuguese utilizing the bigram "intensificação agropecuária" as an element in search of bibliographic databases; additionally Internet search resulted in a significant number of references classified as gray literature. We also tested the use of the unigram "intensification", but it resulted in too large a range of results, suggesting a wide and vague use of the term in our knowledge domain.

We used the software EndNote to perform searches against the Web of Science (ISI) bibliographic database. The result consisted in 1280 references for the period 1964-2009 but we used only a part of this set: 393 references in full text, and 283 in abstracts, composing a textual corpus with 2,570,923 words. After compiling the corpus, we proceeded to the automatic extraction of candidate terms, using the NSP (Ngrams Statistic Package) [11], a set of programs designed to identify and extract n-grams (a contiguous sequence of words) from the corpus, with pre-established parameters. The bibliographical sample proved to be representative because "agricultural intensification" was the most frequent bigram between the total of words extracted from the corpus.

As a second step, we matched the list of candidate terms (extracted from our corpus) with AGROVOC vocabulary <http://aims.fao.org/standards/agrovoc/functionalities/download>, in February 2010. To this end, we developed a tool to compare terms automatically and indicate whether or not they are present in AGROVOC. In case the English word is found in AGROVOC, the tool extracts its translation in Portuguese. The matching gave us the following important outputs: (1) whether a given term already existed or not as a record in the thesaurus; (2) options to choose the terms better suited to represent the concepts and (3) whether a given term was already translated to Portuguese and if such a translation was appropriate from the point of view of Brazilian written and spoken Portuguese.

As a third and final step, we performed a categorization of the terms/concepts to hierarchically organize the resulting vocabulary.

3. Results and Discussion

The result of our work is a categorization system and a model of knowledge organization and representation for AI processes in Brazilian soybean and sugar cane production regions. Such a system is composed of 600 terms or concepts in English, with its corresponding translation to Brazilian Portuguese; around 50% of both, concepts or terms, not yet recorded in AGROVOC, neither in English nor in Portuguese.

The model was organized in four main conceptual levels: “environment”, “agronomy”, “territoriality” and “socioeconomy”. It means that global knowledge about AI results essentially from the understanding of the signification and interrelatedness of concepts from these four knowledge subdomains and their interfaces. The model also includes three other categories that add value to the understanding of AI: “methodologies” utilized in studies of this process; “geographic locations” where these processes are occurring and the “institutions” that are engaged with this subject from RD&I, financial, governmental or not point of views. Table 1, shows the conceptual and hierarchical arrangements of the subdomain "environment", presented in folder tree visualization. The table also highlights (second column) whether the term included in the model was present in AGROVOC or not. The whole model can be found at: <http://cnptia.embrapa.br/~leandro/intagro>, where hierarchical and associative relationships can be seen in graph visualization.

The resulting model is actually open ended, because as the system enters into use, new concepts and terms will be aggregated to it. So, the figures and concepts presented here only reflect the current stage of development. However, we believe that the general conceptual structure is going to remain because we believe that the elements needed to make a good understanding of agricultural intensification process could be gathered and reorganized designing a more complete conceptual model. In this aspect, AGROVOC was very useful because it was possible to recover from this thesaurus several other concepts or terms not extracted from the corpus, but necessary to compose a coherent and representative terminology.

From a conceptual standpoint, the modeling exercise presented here allowed us to propose an interesting scheme representing the multifaceted aspect and multidisciplinary nature of AI processes and indicating that their understanding should consider the interaction and integration of different perspectives of environmental, agronomic, territorial and socioeconomic variables and also considering the need for an analysis of such variables into the context of appropriate methodologies and in specific institutional contexts. AI may in fact have both positive

connotations, when it represents a process of regional progress and development, or negative consequences when, for example, the process is not locally compatible with current concepts of sustainability. Considering its simpler notion, as already mentioned above, i.e., major agricultural production/productivity in the same area at some cost in labor or capital inputs, AI positive aspects can represent, for example, more jobs and capital incomes and, consequently, major socioeconomic profits. On the other hand, if technological improvement is not suitable, we can produce environmental negative impacts as soil fertility losses which in some circumstances can be highly related to AI.

From terminological standpoint this work has demonstrated that AGROVOC was a useful reference for constructing conceptualization and categorization models of agricultural knowledge subdomains. As a counterpart, this work may also contribute to the enrichment of this thesaurus, recovering candidate terms from the literature by textual corpora construction.

Table 1. Fragment of the resulting knowledge organization system for agricultural intensification processes in Brazil. The second column indicates whether terms were present in AGROVOC.

Proposed terminology and categorization system	AGROVOC registered term
*** agricultural intensification (intensificação agropecuária)**	no
(...)	
o environment (meio ambiente)	yes
+ biodiversity (biodiversidade)	yes
# species composition (composição de espécies)	no
* species richness (riqueza de espécies)	no
o Zoology (Zoologia)	yes
+ Insecta (Insecta)	yes
# Hymenoptera (Hymenoptera)	yes
* bumble bee (mamangava[1]; mamangaba[1])	yes
o Botanics (Botânica)	yes
+ plant community (comunidade vegetal; fitocomunidade[1])	yes
# plant species richness (riqueza de espécies [de plantas][2];[vegetais][2])	no
* grass species (espécies de gramíneas)	no
# plant biomass (biomassa vegetal)	no
o microorganisms (micro-organismos)	yes
+ microbial biomass (biomassa microbiana)	no
+ environmental impacts (impactos ambientais)	yes
# degradation processes (processos de degradação)	no
* environmental degradation (degradação ambiental)	yes
o land degradation (degradação das terras)	yes
# ecosystem (ecossistema)	yes
* agricultural ecosystems (ecossistemas agrícolas)	no
o agroecosystems (agroecossistemas)	yes
* community structure (estrutura da comunidade)	no
o food chain (rede alimentar; cadeia alimentar[1])	yes
(...)	

[1]PT/BR term translation not yet present in AGROVOC; [2]Alternative ways to PT/BR term translation; the symbols o, +, #, * represent different hierarchical levels in the categorization system.

4. How about Idiomatic Variations?

AGROVOC vocabularies are presented in English with translations into 21 other languages and Portuguese is one of them. Until now the custody of Portuguese language in AGROVOC was carried out by Portugal native professionals and vocabularies firstly represent the agricultural realities from this country. Portuguese is also the official language in Brazil, but Brazilian Portuguese (PT/BR) obviously presents many idiomatic variations with respect to the Portuguese written or spoken in Portugal (PT/PT).

Historically, in many Brazilian regions, the Portuguese language was firstly influenced by both a multitude of native indigenous languages and by languages of African people brought to Brazil as slaves for a period of more than three centuries. Later, PT/BR was also strongly influenced by

European or Asiatic languages of immigrant people including Italians, Germans, Japanese and Arabians that arrived to Brazil in massive movements in the late nineteenth or early twentieth centuries and settled in the country also developing agricultural or livestock practices.

Moreover, the agriculture practiced in Brazil is immersed in the tropical and subtropical nature of most of its territory, and thus very different from the agriculture practiced in Portugal. Consequently, different agricultural practices, and different concepts and vocabularies are derived. In fact, this situation is not specific to Portuguese, but also holds in the case of others languages widely spoken in the world, like English, French and Spanish.

Currently, the computational design and structure of AGROVOC does not allow for the identification of geographical (national/regional) variations within the same language. For example, if a very specific term of PT/BR is registered in this thesaurus there is no way to recognize it as being specifically used in the Brazilian context. Discussions are taking place within the AGROVOC team, to allow for the identification of such variations, so that users of this terminological resource can identify which region, cultural context and variant idiomatic the term came from.

The terminology referring to agricultural intensification developed in this study revealed several types of idiomatic variations when PT/BR is compared with PT/PT. Examples can be seen in Table 2. They are given to help develop and establish standards to include these variations in AGROVOC.

Table 2. Examples of idiomatic variations between Portuguese from Portugal and from Brazil.

ENGLISH	PT/PT	PT/BR	DIFFERENCES DUE TO	OBSERVATIONS
Agricultural planning	Planeamento agrícola	Planejamento agrícola	Orthography	"planejamento instead of "planeamento"
Reproduction control	Controlo da reprodução	Controle da reprodução	Orthography	"controle" instead of "controlo"
Sterile insect release	Libertação de insectos estéreis	Liberação de insetos estéreis	Orthography	"liberação" instead of "libertação"
Canopy	Coberto arbóreo ou arbustivo	Cobertura arbórea ou arbustiva	Orthography	"cobertura" instead of "coberto"
Land clearance	Rocadura	Roçadura	Orthography	PT/PT and PT/BR must utilize the caracter "ç"; there is no meaning for the word written with caracter "c"
agricultural research	Investigação agrária	Pesquisa agrícola	Other term used in PT/BR	-
Poultry farming	Criação de aves de capoeira	avicultura	Other term used in PT/BR	-
bare fallow	Pousio inculto	Pousio de/com solo nu; pousio nu	Other term used in PT/BR	-
Tillage	Mobilização do solo	Trabalho do solo	Other term used in PT/BR	-
Farm management	Administração exploração agrícola	Gestão agrícola	Other term used in PT/BR	-
Irrigation rates	Dotação de irrigação	Taxa/grau de irrigação	Other term used in PT/BR	-
Sole cropping	Cultivo estreme	Cultivo puro	Other term used in PT/BR	-
Dung	Bosta	Esterco; estrume	Other term used in PT/BR	-
Plot size	Dimensão do talhão	Dimensão da parcela	Other term used in PT/BR	-
Fish processing	Tratamento do pescado	Processamento de pescado	Other term used in PT/BR	-
Food shortages	Penúria alimentar	Escassez alimentar	Other term used in PT/BR	-
Conservation tillage	Mobilização solo para conservação	Práticas de conservação do solo	Other term used in PT/BR	-
Biological control agents	Agente de luta biológica	Agente de controle biológico	Other term used in PT/BR	"controle" instead of "luta"
Weeding	Monda	Capina	Other term used in PT/BR	-
bumble bees	Abelhão	Mamangava; mamangaba	Special case: term from regional linguistic	Brazilian indigenous term
burn tillage *		coivara	Special case: term from regional linguistic	Brazilian indigenous term

PT/PT: Portuguese from Portugal; PT/BR: Portuguese from Brazil; *: term not registered in AGROVOC.

4. Conclusions

In this paper, we presented our work on constructing a knowledge model for “agricultural intensification” (AI). We compiled an English textual corpus with the methodology exposed in Sec. 3, then compared the keywords extracted from the corpus, with the concepts and terms present in AGROVOC. The result was an open ended categorization system, currently with around 600 concepts or terms representing the AI processes observed in soybean and sugar cane Brazilian regions.

AGROVOC proved to be a helpful resource to be used as main element of information and knowledge organization of agricultural domain, as well as a reference for creation of derived terminological products. The general framework of this thesaurus helped us to take the terms and concepts and to prepare the suitable arrangements to representing the considered knowledge domain indicating their hierarchical or associative relationships, synonymies, homonyms, variants, equivalents and polysemies besides, when available, their translation from English into Portuguese language.

The results presented here have also shown us that more work is needed for agricultural domain knowledge organization, representation and mapping. In particular, our work showed the need for expanding AGROVOC to better cover the concepts and terminologies used in Brazil. More in general, our work contributed to make evident that a more general way to represent regional variations of languages in vocabularies such as AGROVOC is needed.

Finally, we expect that the model reported here may facilitate the relationship of the agricultural knowledge generated by Embrapa with those generated by other institutions. Such shared knowledge models should help to properly and effectively connect and retrieve data and information from different organizations, and in so doing, support the formulation of development strategies and policies to improve the sustainability and competitiveness of the Brazilian agricultural sector.

References

1. Boserup, E.:The conditions of agricultural growth.Aldine, New York(1965).
2. Lambin, E.F., Rounsevell, D.A., Geist,H.:Are agricultural land-use models able to predict changes in land-use intensity? Agriculture, Ecosystems and Environment,82, 321-331(2000).
3. Contini, E., Martha, G.: Desempenho da Agricultura Brasileira em 2010 e Perspectivas para 2011, http://blog.elisiocontini.com/wp-content/uploads/blog.elisiocontini.com/2011/02/ContiniGBMJ_Jornal-RS_jan11.pdf
4. Schutte, C., Du Preez, N.: Knowledge networks for managing innovation projects. In: Proceedings of the Portland International Center for Management of Engineering and Technology - PICMET '08, pp. 529-545 (2008), http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04599662.
5. Torres, T.Z., Pierozzi, I., Jr., Bernardes, R. M., Vacari, I.:Collaborative environments in RD&I institutions of the brazilian agricultural sector. J.Technol. Manag. Innov., 5, 69-70 (2010).
6. Torres, T.Z., Pierozz, I., Jr., Pereira, N. R., Castro,A.: Knowledge management and communication in Brazilian agricultural research: an integrated procedural approach. Int. J. Inf. Manage., 31, 121-127 (2011).
7. Bufren, L.S.; Gabriel, R. F., Jr.: A apropriação do conceito como objeto na literatura periódica científica em ciência da informação. Inf. Inf., 16, 52-91 (2011).
8. Zeng, M.L. Knowledge organization systems (KOS). Knowledge Organization, 35, 160-182, (2008).
9. Pierozzi, I., Jr., Oliveira, L.H.M., Souza, K.X.S.:Construindo ontologias de domínio: o (re)conhecimento da intensificação agropecuária no Brasil. In: 3rd Seminário de Pesquisa em Ontologia no Brasil, Universidade Federal Fluminense, Niterói, pp. 100-109 (2010).
10. Fellipo, A.D., Aluísio, S.M., Oliveira, L.H.M., Almeida, G.M.B.: OntoMethodus - a methodology to build domain-specific ontologies and its use in a system to support the generation of terminographic products.In: 6th Workshop em Tecnologia da Informação e da Linguagem Humana - TIL, Sociedade Brasileira de Computação, Porto Alegre, pp. 393-395 (2008).
11. Banerjee, S., Pedersen, T.: The Design, Implementation, and Use of the Ngram Statistics Package. In: Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics, Springer, New York, pp. 370-381 (2008).