Embrapa's contributions to integrate brazilian agricultural vocabularies: Agrotermos in AgroPortal

Milena Ambrosio **Telles**[1,*], Clement **Jonquet**[2], Bibiana Teixeira de **Almeida**[1], Jaudete **Daltio**[1], Celina Maki **Takemura**[1], Leandro Henrique Mendonça de **Oliveira**[1] and Maria de Cléofas Faggion **Alencar**[1]

[1]*Empresa Brasileira de Pesquisa Agropecuária (Embrapa), Brazil*

[2]*MISTEA, Institut Agro, INRAE, University of Montpellier, France; LIRMM, CNRS, University of Montpellier, France*

Abstract

The paper decribes the motivation, challenges and preliminary results of changes made to Agrotermos, Embrapa's conceptual space, to turn it into a semantic resource encoded in standard Knowledge Organization Systems (KOS) formats in order to facilitate it's manteinance and edition, promote its interoperability with other resources and improve the access and visibility to Brazilian Portuguese semantic resources and henceforth to Embrapa's results in Brazilian research in Agriculture and related domains. These changes were made within the scope of a partnership established with AgroPortal, a web-based platform designed to support data integration, sharing, and analysis, and which offers tools and services to manage and use ontologies and semantic resources. Agrotermos' infrastructure and processes are being reformulated, VocBench is being implemented as its data management tool, and its contents are being prepared and validated for publication in AgroPortal. The preliminary results of the changes to Agrotermos' infrastructure show that both VocBench and AgroPortal largely improve editing, visualization of the semantic content and validation of mappings and alignments between vocabularies.

Keywords

semantic resources, knowledge organization systems (KOS), interoperability, terminology

1. Introduction and Motivation

Human knowledge is structured using mostly written languages, and human languages recorded in the form of texts produce immense sets of data and information. Organizations involved in Research, Development and Innovation, such as the Brazilian Agricultural Research Corporation (Embrapa) or France's National Research Institute for Agriculture, Food and Environment (INRAE), face challenges to process all their data to extract knowledge for decision support [1].

Knowledge Organization Systems (KOS) that model the semantic structures of specialized domains become essencial to process, organize, systematize and manage such data and information, and hence foster innovation [2]. These KOS, materialized as semantic resources/artefacts (such as ontologies, terminologies and thesauri), enable identifying standards, tendencies and insights off complex datasets and provide valuable support to strategic decisions [3].

Embrapa was established to lay technical and technological foundations for tropical agriculture and animal farming, and is one of the largest agricultural research corporations in the world [4]. The company's technologies and solutions are produced and offered mostly in Brazilian Portuguese, but have good outreach potential to address the whole tropical world. Aside from Brazil, Portuguese is an official language in eight other countries in four continents and is spoken by over 260 million people worlwide.

With the aim of addressing some of this challenges using terminology to add value to and expand the company's knowledge management expertise, capacity and outreach, Embrapa created Agrotermos, a conceptual space for knowledge representation of agriculture and its related areas. Operational constraints limited the use of this semantic resource to its full potential, which motivated a thorough study of improvements and updates necessary to both Agrotermos' infrastructure and processes. The goal of this paper is to present an overview of the challenges faced and the technological and procedural strategy currently put in place to make Agrotermos more operational and available to its users. This article is divided into three parts and the introduction, which describe 1) the scenario of the ongoing work, a brief description of the objects involved and the main challenges encountered; 2) the results achieved so far; 3) the preliminary conclusions and the next steps to be followed.

2. Research Scenario

2.1. Agrotermos

Since 2018, Embrapa has a designated Permanent Commission for Controlled Vocabularies, Agriterminologies and Agrisemantics (GTermos), which is committed to assembling, sharing, disseminating and supporting terminological and semiotic support to knowledge, data and information management initiatives by Embrapa and its partners. Its actions are aimed at increasingly aligning Agrotermos with global standardisation initiatives, such as W3C, and tendencies, such as the FAIR Principles (Findable, Accessible, Interoperable, Reusable) for knowledge and data information management processes within the company. GTermos, a team of ten people, is responsible for managing and editing Agrotermos.

Agrotermos (https://sistemas.sede.embrapa.br/agrotermos/) was conceived to aggregate the main controlled vocabularies on agriculture and related areas available in Portuguese [5]. Built based on the method and technology proposed by the Global Agricultural Concept Space (GACS) [6], Agrotermos contains the full Brazilian Thesaurus Agrícola Nacional (https://sistemas.agricultura.gov.br/tematres/vocab/index.php - Thesagro) and the Portuguese content of FAO's (United Nation's Food and Agriculture Organization) Agrovoc Multilingual Thesaurus, as well as other resources created by Embrapa, along with alignments and mappings created by the GTermos team between these resources. Agrotermos comprises a platform for organizing, qualifying, and offering terminology data and semantic applications produced within the company.

Agrotermos' version 1.0, built in 2014, contains nearly 56 thousand terms, and merges in one unique conceptual space the content from multiple sources to use them all uniformly in an integrated approach. It may presently be accessed twofold: (1) by a web interface, intended for human use; and (2) by an application programming interface (API) that uses web-service Representational State Transfer (REST) technology, for machine-machine communication. These interfaces enable searching and browsing the semantic resource's content (concepts and their relations) as served from a relational database. This version of Agrotermos has been accessible only via the custom eponym application, but did not exist as an independent source file (typically using the SKOS language representation and the corresponding RDF syntaxes) that could be distributed to users and explored by multiple tools and products using semantic web standards.

However, since GTermos is also part of the Agrovoc editorial community as the editor of Brazilian Portuguese[7], Agrotermos' curation models and practices have drifted increasingly closer to those of Agrovoc, and hence to relevant international standards, processes, resources and tools. Over time, challenges and requirements have arisen, and advancements to Agrotermos became necessary, including the need for GTermos/Embrapa to be able to distribute Agrotermos, either whole or in parts, as a semantic resource encoded in a standard format such as SKOS. In 2024, ten years into Agrotermos' creation, new taskforces have been formed by GTermos, and its partners within and outside Embrapa, to address these issues and to select and implement up-to-date, standard-aligned procedures and infrastructures.

2.2. AgroPortal

AgroPortal (https://agroportal.lirmm.fr/) is a web-based platform that provides access to a wide range of semantic resources and ontologies in the agri-food domain [8]. It is designed to support data integration, sharing, and analysis by offering tools and services to manage and use ontologies and semantic resources effectively. The platform is based on the OntoPortal technology [9] and features ontology hosting, search, versioning, visualization, comment, recommendation, enables semantic annotation, as well as storing and exploiting ontology alignments, all within a fully semantic web compliant infrastructure. It currently hosts 188 semantic resources in different formats, and languages.

2.3. Research Challenges

The main challenges faced by Agrotermos are twofold: 1. Upgrading its technological infrastructure to one that is sourced by and operates under up-to-date semantic standards, and 2. Offering a visualization (distribution) interface that enables different users (librarians, text editors, terminologists, etc.) to easily access and use its data.

Regarding Agrotermos' technological infrastructure, the main challenge was mapping and addressing all the main technological and procedural aspects required to transform it into a fully interoperable semantic resource that meets both the international standards for resources of its kind and the FAIR Principles. This means that the set of technologies adopted should be a standard-compliant, easy-to-use, web-accessible platform for the curation and editing of different kinds of semantic resources. It should feature different user profiles, to enable hierarchical decisions and reviews, as well as different output formats. This "vocabulary management tool" should enable the operation of semantic resources in Portuguese (Brazilian Portuguese) and English, as well as the creation of mappings and alignments between its own and other semantic resources, e.g., controlled vocabularies with either poly- or monohierarchical structure. Furthermore, its outputs should be seamlessly interoperable with other semantic resources. This required revisiting, exercising and redefining the processes involved in editing and inserting new semantic resources into Agrotermos.

To address this issues, and in partnership with AgroPortal, GTermos decided to embrace SKOS natively to distribute the content of Agrotermos and then rely on the vocabulary platform to distribute, serve and visualize this content. Between SKOS and OWL, which are the two W3C recommendations for representing semantic resources, SKOS was the most appropriate. It provides the means and constructs appropriate for the level of semantics needed for Thesagro and VocGeo, which have no need for the complex semantic layer of OWL. It is also the representation language already adopted by Agrovoc.

This provides greater visibility for Agrotermos' data and, consequently, for Brazilian agriculture. In AgroPortal, new knowledge insights are possible between Agrotermos and other ontologies and semantic resources in the repository. Plus, this choice is coherent with the fact that Agrovoc, partly included in Agrotermos, is itself hosted and served by AgroPortal.

Nevertheless, adopting SKOS and using AgroPortal to distribute Agrotermos' resources would address only the requirements related to the worldwide standard "distribution" of its results. There remained a need to adopt new standard tools and methods to "produce" the resources and use a standard vocabulary edition software such as VocBench[10].

3. Results

We have rethought the whole Agrotermos original infrastructure to address the aforementioned issues. The use of more structured technologies for managing, visualizing, and interoperating Agrotermos' terminological data and semantic resources required embracing rather thorough technological advancements. Figure 1 shows the proposed new structure.

The new structure equips the GTermos team with the multilingual platform VocBench to manage, edit and produce SKOS distributions of the resources contained in Agrotermos. The choice for VocBench as the data management tool for Agrotermos was based on the team's own experience in editing the
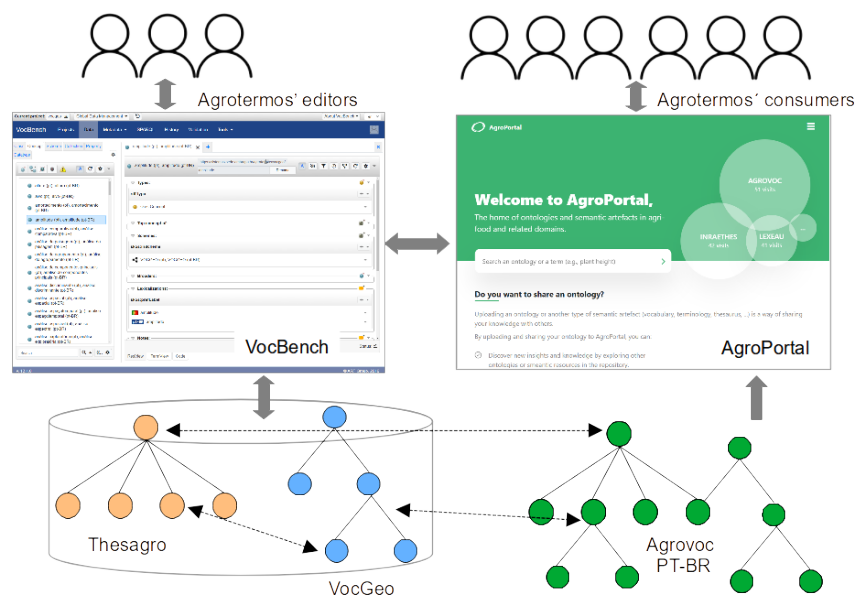
Figure 1: Overview of Agrotermos' proposed new structure.

Brazilian Portuguese content in Agrovoc. VocBench is a web-based, multilingual, collaborative platform for managing OWL ontologies, SKOS thesauri, Ontolex-lemon lexicons, generic RDF datasets, and other semantic resources.

Data from Thesagro and other terminological datasets, controlled vocabularies, and/or ontologies generated by Embrapa and its partners may be incorporated into the new Agrotermos environment. Portuguese-language Agrovoc data were part of the original Agrotermos database. However, since Agrovoc is now fully integrated into AgroPortal, and all the language editing for Agrovoc occurs within its own VocBench project, there is no longer a need to replicate its content in Agrotermos' VocBench. In the new proposed structure, AgroPortal will not only display Agrotermos' data in Brazilian Portuguese, but also align its content with all other semantic resources hosted on the portal, including Agrovoc [8].

To enable the implementation of the new tools and the establishment of new management processes for Agrotermos, four workgroups were assembled to:

1. adapt Thesagro, itself included in Agrotermos, directly from its original BINAGRI Tematres installation to SKOS and include it in AgroPortal;

2. adapt VocGeo, a vocabulary on geoinformation produced by Embrapa (also included in the Agrotermos database), to SKOS and include it in AgroPortal;

3. study and install VocBench as the management tool for Agrotermos' semantic resources and mappings;

4. organise a group of semantic resources in AgroPortal called Agrotermos, which contains Agrovoc, Thesagro and VocGeo for the moment, but may comprise other Embrapa's semantic resources in the future.

The data adaptation mentioned in items 1. and 2. included generating the complete SKOS format and uploading it to Agroportal.

Thesagro and VocGeo, as well as the alignments and mappings that existed between them and Agrovoc's content in Portuguese (PT) and Brazilian Portuguese (PT-BR), have already been inserted into AgroPortal. Thesagro and VocGeo are in their final validation phase to become publicly available without restrictions, and Agrovoc is already fully available. AgroPortal regrouped theses three resources within the Agrotermos group, and a slice was created for Agrotermos, to allow users to search and browse an exclusive AgroPortal view of only these three semantic resources: https://agrotermos.agroportal.lirmm.fr/.

4. Conclusion and Ongoing Work

The first results of the changes to Agrotermos' infrastructure show that both VocBench and AgroPortal largely improve both editing, the visualization of the data and the validation of mappings and alignments between vocabularies. Furthermore, once VocGeo and Thesagro are fully, publicly available, they will get a FAIR score produced by O'FAIRe tool in AgroPortal.

There is still a need for training the GTermos team, which is at the beginning of its learning curve to master the use of the new tools to their full potential.

Furthermore, the new proposed infrastructure has drawn GTermos and Thesagro's team to negotiations to collaborate in joint infrastructure and processes for editing both Agrotermos and Thesagro.

Acknowledgments

References

[1] I. Pierozzi Junior, P. R. B. Bertin, C. de Laia Machado, A. R. da Silva, Towards semantic knowledge maps applications: modelling the ontological nature of data and information governance in a RD organization, InTech, Rijeka, 2018. doi:10.5772/67978.

[2] M. L. Zeng, Knowledge organization systems (KOS), Knowledge Organization 35 (2008) 160–182. doi:10.5771/0943-7444-2008-2-3-160.

[3] C. G. Duque, G. G. Bastos, Ontologia aplicada a um modelo de gestão organizacional: contribuições da ciência da informação, Ciência da Informação 46 (2017) 197–213. doi:10.18225/ci.inf.v46i1.4023.

[4] E. S. de Comunicação., Seu futuro inspira a nossa ciência, Brasília, DF, 2023.

[5] Embrapa, Agrotermos, https://sistemas.sede.embrapa.br/agrotermos/, 2014.

[6] I. Pierozzi Junior, M. C. Visoli, M. I. F. Souza, L. M. S. Cunha, I. Vacari, T. Z. Torres, Engenharia da informação: contribuições para a agricultura digital, Embrapa, 2020, pp. 192–217. URL: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/217705/1/LV-Agricultura-digital-2020-cap8.pdf.

[7] FAO, Meet the agrovoc editorial community, https://www.fao.org/agrovoc/news/meet-agrovoc-editorial-community-1, 2021.

[8] C. Jonquet, A. Toulet, E. Arnaud, S. Aubin, E. Dzalé Yeumo, V. Emonet, J. Graybeal, M.-A. Laporte, M. A. Musen, V. Pesce, P. Larmande, Agroportal: A vocabulary and ontology repository for agronomy, Computers and Electronics in Agriculture 144 (2018) 126–143. URL: https://www.sciencedirect.com/science/article/pii/S0168169916309541. doi:10.1016/j.compag.2017.10.012.

[9] C. Jonquet, J. Graybeal, S. Bouazzouni, M. Dorf, N. Fiore, X. Kechagioglou, T. Redmond, I. Rosati, A. Skrenchuk, J. L. Vendetti, M. Musen, Ontology Repositories and Semantic Artefact Catalogues with the OntoPortal Technology, in: ISWC 2023 - 22nd International Semantic Web Conference, volume 14266 of *Lecture Notes in Computer Science*, Springer, Athens, Greece, 2023, pp. 38–58. URL: https://hal.science/hal-04088537. doi:10.1007/978-3-031-47243-5_3, members of the OntoPortal Alliance.

[10] A. Stellato, M. Fiorelli, A. Turbati, T. Lorenzetti, W. Gemert, D. Dechandon, C. Laaboudi-Spoiden, A. Gerencsér, A. Waniart, E. Costetchi, J. Keizer, Vocbench 3: A collaborative semantic web editor for ontologies, thesauri and lexicons, Semantic Web 11 (2020) 855–881. doi:10.3233/SW-200370.