Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 32, Num. 3 (2025) 102-111

RESEARCH ARTICLE

An Ontology of Tobacco Production: Enriching Large Language Model-based Decision Support

Ontologia da Produção de Tabaco: enriquecendo o suporte à decisão baseado em LLM

Luis Felipe Medeiro Alves[1]*, José Maria Parente de Oliveira[2], Rodrigo Bonacin[3], Ferrucio de Franco Rosa[2,3]

Abstract: Tobacco (*Nicotiana tabacum*) production plays a crucial role in the agricultural economy of several regions around the world, especially in developing countries, such as Brazil. However, the lack of well-defined terminology and semantic models harms the development of decision support systems. Conceptualizing the tobacco production domain is challenging due to its ambiguous terminology and the complexity involved in considering environmental, soil, disease, and pest management factors. We present an Ontology of Tobacco Production (OnTop), designed to assist tobacco production in optimizing crop management practices and placing environmental factors within a formal framework. To the best of available knowledge, this study is the first to formalize the tobacco production lifecycle integrated with soil and climate data. It includes symbolic and description logic for reasoning and automation. This paper presents the proposal for the core set (main classes) and the application of OnTop to enrich a Large Language Model (LLM) to provide an ontology-based decision support prompt. OnTop allows actionable recommendations based on environmental, agronomic, and productivity data. The main contributions of this paper are: i) a domain ontology that formalizes knowledge on data-driven tobacco production; and ii) a queryable framework (enriched LLM), which allows experts to obtain complex agronomic answers. OnTop offers an extensible framework for decision making in tobacco farming and paves the way for further innovations in ontologies and LLM-based decision support systems for the agricultural domain.
Keywords: ontology — knowledge representation — tobacco — large language model

Resumo: A produção de tabaco (*Nicotiana tabacum*) desempenha um papel crucial na economia agrícola de diversas regiões do mundo, especialmente em países em desenvolvimento, como o Brasil. No entanto, a ausência de uma terminologia bem definida e de modelos semânticos prejudica o desenvolvimento de sistemas de apoio à decisão. A conceituação do domínio da produção de tabaco apresenta desafios devido à ambiguidade terminológica e à complexidade inerente à consideração de fatores ambientais, edáficos, fitossanitários e de manejo de pragas. Apresentamos a Ontologia da Produção de Tabaco (OnTop), desenvolvida com o objetivo de auxiliar na otimização das práticas de manejo agrícola e na formalização de fatores ambientais por meio de um arcabouço ontológico. Até onde se tem conhecimento, este estudo é o primeiro a formalizar o ciclo de vida da produção de tabaco de forma integrada a dados de solo e clima, incorporando lógica simbólica e descritiva para fins de raciocínio automatizado. Este artigo apresenta a proposta do núcleo conceitual (classes principais) e a aplicação da OnTop no enriquecimento de um Modelo de Linguagem de Grande Escala (LLM), de modo a viabilizar a geração de prompts baseados em ontologias para apoio à decisão. A OnTop possibilita a emissão de recomendações acionáveis com base em dados ambientais, agronômicos e de produtividade. As principais contribuições deste trabalho são: i) uma ontologia de domínio que formaliza o conhecimento orientado por dados sobre a produção de tabaco; e ii) um arcabouço consultável (LLM enriquecido), que permite a especialistas obter respostas agronômicas complexas. A OnTop constitui um framework extensível para a tomada de decisão na cultura do tabaco e abre caminho para novas inovações no uso de ontologias e de sistemas de apoio à decisão baseados em LLMs no domínio agrícola.
Palavras-Chave: ontologia — representação do conhecimento — tabaco — modelo de linguagem de grande escala

[1] *University of São Paulo (USP), Brazil*
[2] *Aeronautics Institute of Technology (ITA), Brazil*
[3] *Renato Archer Information Technology Center (CTI), Brazil*
***Corresponding author**: alves.luis@alumni.usp.br
DOI: http://dx.doi.org/10.22456/2175-2745.146658 • **Received:** 30/03/2025 • **Accepted:** 27/07/2025

1. Introduction

Tobacco *(Nicotiana tabacum)* production plays a key role in the economy of several regions around the world, especially in developing countries such as India, Brazil, Mozambique, Zimbabwe and Indonesia [1]. In Brazil, it is widely cultivated in the southern states. According to the Brazilian Agricultural Parliamentary Front (FPA (2020) [2]), 556 cities in southern Brazil produce tobacco on more than 280 thousand hectares, involving approximately 600 thousand people and generating more than 2 billion dollars in revenue per year. Tobacco represents 46% of the gross income of producers, although it represents only, on average, 19% of the arable area of the properties. For example, a hectare of tobacco generates the same revenue as seven hectares of corn. Brazil exports 85% of the tobacco produced, 40% to the European Union, the United States, and China being other important buyers. Brazil is the second largest tobacco producer worldwide and the leader in exports. As production in the country is mostly restricted to the three southern states, this shows the importance of this culture for the economic and social life of southern Brazil. Although Brazilian production occurs mainly in southern states, reports show that tobacco production in the arid lands of northeast Brazil is increasing at impressive rates [3].

Nicotiana tabacum is the main tobacco crop and the most important commercial non-food crop worldwide [4]. Tobacco farming is labor intensive and, despite the economic importance of this crop, the industry has been slow to adopt technological advances, both for the mechanization of production and decision support systems. Balancing environmental, soil, disease, and pest management factors in the various stages of tobacco production is challenging and requires careful and informed decision making. There is a lack of structured and technology-driven solutions to help farm owners and managers optimize their yields.

We present the Ontology of Tobacco Production (OnTop), designed to assist tobacco production in optimizing crop management practices. OnTop is the first of its kind, as no existing ontologies focus on tobacco production. OnTop formalizes the tobacco production lifecycle, integrates soil and climate data, and improves automated reasoning. The core set proposal (main classes) is described and formalized using the Web Ontology Language (OWL)[1]. Our work delivers a structured and formalized representation of the entire tobacco production lifecycle, from seedbed treatment to post-harvest curing and yield assertion. By formalizing the entities, processes, and relationships that characterize tobacco farming, OnTop aims to provide a decision support tool for farm owners and managers, researchers, government officials, and agricultural advisors.

To integrate OnTop with a large language model (LLM), our objective is to take advantage of the formalized ontology structure to enhance the decision making capabilities of the LLM in tobacco farming. By embedding OnTop into the LLM's knowledge base, the model can utilize predefined classes and relationships to improve the accuracy and depth of its responses. This integration allows the LLM to provide more contextually aware and domain-specific recommendations, such as addressing environmental factors, optimal farming practices, and the management of pests and diseases. Consequently, the enriched LLM can generate actionable and precise answers for complex agronomic queries, moving beyond generic responses to ones tailored by the ontology's semantic framework. Through this enhancement, the system bridges the gap between data-driven insights and human decision making in tobacco production.

This research was guided by the following research question: *"Can a formal ontology of tobacco production, when integrated with a large language model (LLM), improve the accuracy and contextual relevance of agronomic decision support in tobacco farming?"*.

Based on the research question, we formulate the following hypothesis: *"Integrating a domain-specific ontology (OnTop) into a large language model enhances the quality, contextual accuracy, and precision of decision support responses for tobacco production weather compared to a generic LLM without ontology enrichment."* This hypothesis underpins our investigation and the evaluation presented in this work.

The remainder of this article is organized as follows: *Methods* presents the literature review and related work and a discussion of them; *Results and Discussion* presents the core classes of OnTop, an application of OnTop to an LLM and the conclusion.

2. Methods

A literature review [5] was carried out to identify work on agricultural recommendation systems. The review identified the need for work that can systematically support recommendations through artificial intelligence. Ontologies can help with this task. The methodology used in our review of the literature was based on the guidelines proposed by Kitchenham (2004) [6].

2.1 Background and Motivation

Today, advances in agricultural technologies, particularly those that involve Geographic Information Systems (GIS) and precision farming tools, have revolutionized how farmers manage their crops. GIS technologies can improve yields by enabling farmers to optimize their resources, such as water, fertilizers, and pesticides, based on detailed spatial and environmental data [7]. Despite the widespread use of these technologies in various crops, tobacco production has been relatively slow to adopt such innovations, especially in developing countries. Given the complex nature of tobacco cultivation, which involves a delicate balance of climate, soil, and pest and disease management, there is significant potential to improve yields through data-driven decision support systems.

[1]The Web Ontology Language (OWL) is a Semantic Web language designed to represent complex knowledge. More details are available on the W3C website: https://www.w3.org/OWL/

Our study [5] builds on previous research in GIS and agricultural decision support systems by integrating these concepts with an ontology-based reasoning framework for the specific reality of tobacco production. The primary advantage of an ontology is its ability to capture complex relationships between entities in a structured and formal manner. This formalization allows automated reasoning, enabling the ontology to provide recommendations to farmers based on historical and real-time data. Querying an ontology for specific recommendations can improve decision-making, positively impacting yield in tobacco crops, where environmental factors and human interventions influence different stages of plant growth.

In the context of developing agricultural Decision Support Systems (DSS) and Geographic Information Systems (GIS)-based applications, recent works have explored the integration of GIS into Recommender Systems (RS) to optimize agricultural practices, such as [8][9][10][11][12]. In our study, we identified methods and technologies that aim to offer recommendations on soil management, crop selection, fertilization, pest control, and other critical agricultural activities, such as [13][14][15][16][17][18].

The motivation for developing an ontology for decision support systems (DSS) in the context of tobacco production stems from the growing need for more robust and specific solutions in agricultural management. The use of ontologies in DSS has shown potential in other agricultural domains by formalizing complex relationships between environmental, biological, and management data, allowing more informed and context-aware decision-making [19]. In the case of tobacco, where the balance between factors such as climate, soil, and pest control is critical, the formalization of an ontology such as OnTop provides a solid foundation for automating and improving recommendations. This is achieved by integrating real-time and historical data from geographic information systems (GIS) and precision agriculture tools.

2.2 Related Work

We focus on presenting related works on ontologies for tobacco. Although ontologies have been developed for other crops and domains, the tobacco industry has lacked this level of formal knowledge representation so far. The current tobacco-related ontologies focus on addiction and health (e.g. [20] [21]), however, for the agricultural production of tobacco, in terms of ontologies, nothing had been found in our review of the literature [5].

Cox et al. (2023) [20] focus on developing an ontology that formalizes identity-related constructs in addiction research, focusing on nicotine and tobacco. The main objective of their work is to create a framework for identity constructs that can be applied across various addiction theories, facilitating a common language to be used by researchers. Their ontology defines key concepts like self-identity, group identity, and the cognitive representations that individuals or groups hold about themselves, such as ”tobacco smoker identity” or ”vaper identity.” In addition, they distinguish between identities that individuals positively or negatively appraise, offering a framework that accounts for how these identities are valued and maintained over time. This focus on the psychological and social aspects of tobacco use, addiction, and identity formation fills a gap in addiction research by allowing a more structured representation of these constructs. However, while Cox et al. contribute to understanding nicotine addiction and identity in the context of tobacco use, their work does not extend to the agricultural or environmental aspects of tobacco production. Their ontology is not designed to assist in decision making related to tobacco farming, as it focuses on behavioral and psychological dimensions.

Notley et al. (2022) [21] focus on identity-related constructs in addiction, aiming to create a structured ontology that helps clarify the relationships between identity and addiction behaviors, such as smoking and vaping. Their research provides a theoretical foundation for understanding how identity impacts behaviors around nicotine use, proposing distinctions between different forms of self-identity (e.g., ”ex-smoker identity”) and group identity. The ontology also addresses the role of self-apraisal in shaping these identities, introducing terms like ”core identity” for the central and enduring identities. Notley et al.'s work is geared toward improving the clarity of discussions around addiction theories and offering a common language for diverse research approaches in addiction studies. However, like Cox et al., this research is focused on the personal and social dimensions of tobacco use, rather than the agricultural processes or environmental factors associated with tobacco production. While their ontology contributes to addiction research, it is not intended to model the lifecycle of tobacco farming, nor integrate data-driven agronomic insights.

Our work presents an ontology developed for tobacco production. Whereas the studied ontologies deal with psychological and behavioral aspects, OnTop formalizes the entire tobacco production lifecycle, incorporating environmental, soil, and climate data, offering reasoning capabilities to provide a better decision support system for tobacco farmers. This represents a novel application of ontological modeling, as it targets the agricultural management of tobacco rather than its consumption and addiction.

This work aims to contribute to a significant improvement compared to traditional and experience-based decision-making processes. OnTop provides a queryable framework to be used with LLM-based decision support systems. Thus, an ontology for tobacco production is an innovation in agricultural technology.

2.3 Discussion on the Literature Review

Our ontology is intended to structure and represent knowledge throughout the tobacco production stages. Alves et al. (2024) [5] focus on various types of crops and environmental factors, and provide insight into the broader potential of GIS in supporting agricultural decision-making. Specifically, the study highlights the growing importance of integrating spatial and environmental data for optimized agricultural outcomes, which is central to our decision support ontology for tobacco production.

Our ontology builds on these concepts by providing a queryable framework that can be integrated with GIS, allowing farmers to make informed decisions based on real-time environmental data. While in [5] the need for modernization in agricultural practices is emphasized, particularly in developing countries, our ontology extends this by formalizing the complex relationships inherent in tobacco production, contributing to academic discussion and practical advancements in the representation of agricultural knowledge.

3. Results and Discussion

The Ontology of Tobacco Production (OnTop) is a domain ontology planned to formalize the tobacco production lifecycle. OnTop was developed using a formal approach to capture the complexities of tobacco production, allowing users to query for detailed, context-sensitive answers using LLMs.

We used the Basic Formal Ontology (BFO) [22] as a reference to connect our domain proposal to the top-level knowledge representation. BFO is a top-level ontology designed to support information integration, retrieval, and analysis across all domains of scientific investigation. The class **Object** (this class represents continuants) of BFO was chosen as the connection point for the Tobacco Production Domain due to its ability to represent independent entities that persist over time, such as tobacco plants, soil, and environmental factors. These objects are crucial in modeling agricultural systems, as they interact dynamically with other aspects, such as climatic conditions and farming interventions, to influence tobacco production.

3.1 Formalizing the Tobacco Production

We structured our ontology using the following base scenario: experts can ask questions about the optimal timing of fertilization, pest control, or irrigation, and Ontop, or an OnTop-enriched AI tool (e.g., LLM), would respond with recommendations that consider soil conditions, climate predictions, and crop growth stages. Integrating these various factors into a single structured framework improves decision making and allows for future advances in agricultural ontologies, particularly those with regard to tobacco. Figure 1 presents the proposal for the core set (main classes) of OnTop.owl (available in a GitHub repository [23]).

The core structure of OnTop comprises 9 main classes, each representing a fundamental aspect of tobacco production. These classes, which directly inherit from ***TobaccoProduction***, serve as the foundational framework for modeling components and processes in the tobacco farming domain.

The ***ClimateCondition*** class encompasses critical environmental factors such as *Temperature*, *Humidity*, *RainfallAmount*, *SunlightExposure*, *FrostIncidence*, and *HailstoneOccurrence*. These elements influence the growth and health of tobacco plants, affecting decisions on planting times, irrigation, and pest control. Each subclass provides detailed information on specific climate attributes, allowing for precise modeling of environmental impacts on tobacco cultivation.

ManagementAction represents various activities and interventions carried out during tobacco production. This class includes activities such as *FertilizerApplication*, *PesticideApplication*, *Irrigation*, *Weeding*, *SoilPreparation*, and *LimeApplication*. By formalizing these steps, OnTop supports the optimization of agricultural practices, allowing better planning and resource allocation. Structured representation of these management activities also helps to assess their effectiveness in crop yield and quality.

PestsAndDiseases addresses biotic stress factors that can adversely affect tobacco production. It includes subclasses such as *IncidenceOfPestsAndDiseases* and *NoxiousWeeds*, detailing specific threats to the tobacco crop. This classification is crucial for integrating pest and disease management strategies into the broader decision-support framework, ensuring that mitigation measures are timely and effective.

ProductionStage details the 5 phases of the tobacco lifecycle, including ***TobaccoSeedbed***, ***TobaccoPlanting***, ***TobaccoTopping***, ***TobaccoHarvest***, and ***TobaccoCuring***. These stages are sequential and interdependent, considering specific management requirements and environmental considerations. This class and its subclasses aim to facilitate detailed planning and monitoring throughout the production process, ensuring that each stage is executed optimally to maximize yield and quality.

TobaccoLifecycle provides a comprehensive view of the growth and development of tobacco plants, encapsulating entities such as *TobaccoSeed*, *TobaccoSeedling*, *TobaccoPlant*, and *TobaccoLeaf*. This class is essential for modeling the biological and phenomenological aspects of tobacco production, enabling a holistic understanding of the growth dynamics of the crop. It interacts closely with other main classes, such as ***SoilCondition***, illustrating the complex interaction between biological and environmental factors in tobacco farming.

HumanAgent represents people involved in tobacco production, such as farmers, agricultural advisors, and insurance agents. This class is crucial to capture the roles and responsibilities of stakeholders in managing and supporting the tobacco production lifecycle. By formalizing the interactions and decisions of human agents, OnTop enables a better understanding of how human factors influence agricultural results and the implementation of best practices in tobacco farming.

SoilCondition models the soil properties that are critical for tobacco cultivation, including *pH*, *SoilMoisture*, and *SoilType*. These attributes are essential to determine the soil's suitability for tobacco production and for making informed decisions regarding soil management practices (e.g., fertilization and irrigation). Integrating soil conditions into our ontology allows for more precise recommendations for optimizing plant health and yield.

TobaccoField encompasses the spatial and geographical

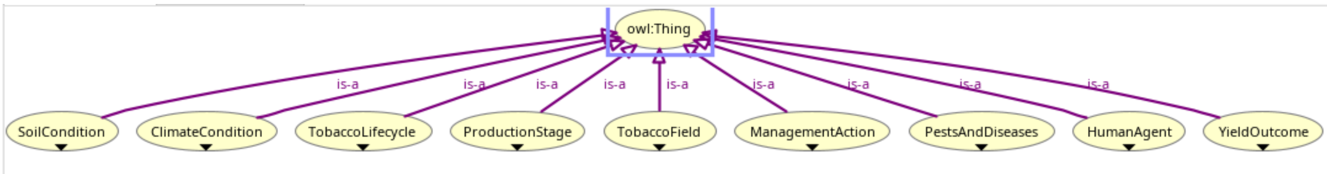


Figure 1. Main Classes of OnTop.owl, showing the hierarchical relationships in the ontology structure.

aspects of tobacco cultivation, including the *GeographicalLocation*, *Size*, and *Soil* of each field. This class is pivotal for linking the physical attributes of tobacco fields with environmental and management data, enabling location-specific recommendations and interventions. By representing each field's unique characteristics, our ontology supports the application of precision agriculture techniques tailored to individual plots.

Finally, the ***YieldOutcome*** class quantifies the results of tobacco production, including metrics such as *YieldPerHectare* and overall *YieldOutcome*. This class is essential to evaluate the effectiveness of various management practices and environmental conditions on tobacco production. By integrating yield data with other ontological elements, our ontology provides a robust framework for assessing the impact of different variables on productive outcomes and guiding future decision-making.

3.1.1 Tobacco Production Lifecycle: an example of the initial phase

The ***TobaccoLifecycle*** class comprises 5 stages with a rigid linear order, each stage has management requirements and is associated with particular environmental factors. Figure 2 presents the tobacco production lifecycle. In our ontology, these stages are represented as subclasses of the ***ProductionStage*** class, with relationships between them captured by object properties. For example, the rigid temporal line is handled by the object properties *precedes* and *follows*. Each stage in the lifecycle is critical to the overall success of the crop, and the ontology models the stages and their influences. The stages can be queried individually or as part of more complex and interrelated queries.

As we did not consider preproduction activities, for example seed acquisition, the first stage of tobacco production is the seedbed (***TobaccoSeedbed*** class). In this stage, the seeds go through treatments (e.g., pelleting, chemical, hormonal, and antibacterial) accounted for by data properties (e.g., *chemicalyTreatedWithBactericide*). The appropriate substrate (has a Boolean data property *isRenewable*) is selected to differentiate nature-friendly substrates (e.g., coconut fiber and pine bark) and non-renewable ones (e.g., peat); it also encompasses the specific sowing process. These activities represent the core of the seedbed stage, and several actions influence them, e.g., fertilizers and defensives may be applied earlier and trimming must occur.

We illustrate this stage by using an example of the formalization of the process in formal symbolic logic. We can represent the seedbed process as follows:

$$S(x) \wedge S_t(x,y) \wedge S_s(x,z) \wedge P_s(x,S(x))$$

Where $S(x)$ represents the seedbed, $S_t(x,y)$ the seed treatment y in the seed x, $S_s(x,z)$ the substrate z for the seed x, and $P_s(x,S(x))$ the seed sowing x in $S(x)$. This represents the relationship between activities in the seedbed phase.

In description logic (DL)[2], we can express the same relations as follows:

$$S \sqsubseteq \exists S_t.\text{SeedTreatment} \sqcap \exists S_s.\text{Substrate} \sqcap \exists P_s.\text{Sowing}$$

This DL expression states that the seedbed phase of the lifecycle involves specific seed treatment, substrate selection and application, and seed sowing on the seedbed trays.

As another example, the equivalent class ***TreatedSeed*** represents a tobacco seed that undergoes some treatment. We can formalize it in the Manchester syntax[3] (used in Protégé[4]) as follows:

```
TobaccoSeed
and ((chemicallyTreatedWithBactericide value
true)
or (chemicallyTreatedWithFungicide value true)
or (chemicallyTreatedWithInsecticide value
true)
or (wasDisinfected value true)
or (wasHormonallyTreated value true)
or (wasPelleted value true))
```

It can also be modeled formally in symbolic logic as follows:

$$\forall x \Big(T(x) \wedge \big(C_B(x) \vee C_F(x) \vee C_I(x) \vee D(x) \vee H(x) \vee P(x)\big) \rightarrow TreatedSeed(x)\Big)$$

Where:

- x: x is an individual of the ***TobaccoSeed*** class.
- $T(x)$: x is a tobacco seed.

[2]Description logics are knowledge representation languages, used in conceptual or ontological modeling.

[3]The Manchester syntax is a compact syntax for OWL 2 ontologies. Link: https://www.w3.org/TR/owl2-manchester-syntax/

[4]Protégé is a free, open-source ontology editor. Link: https://protege.stanford.edu/

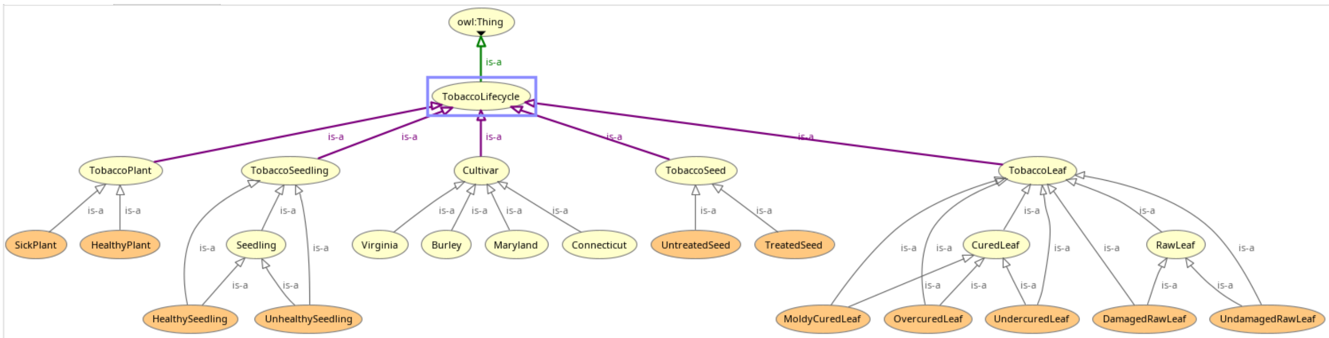


Figure 2. Tobacco Production Lifecycle.

- $C_B(x)$: x is chemically treated with a bactericide.
- $C_F(x)$: x is chemically treated with a fungicide.
- $C_I(x)$: x is chemically treated with an insecticide.
- $D(x)$: x was disinfected.
- $H(x)$: x was hormonally treated.
- $P(x)$: x was pelleted.
- $TreatedSeed(x)$: x is a treated seed.

And, similarly, in DL as:

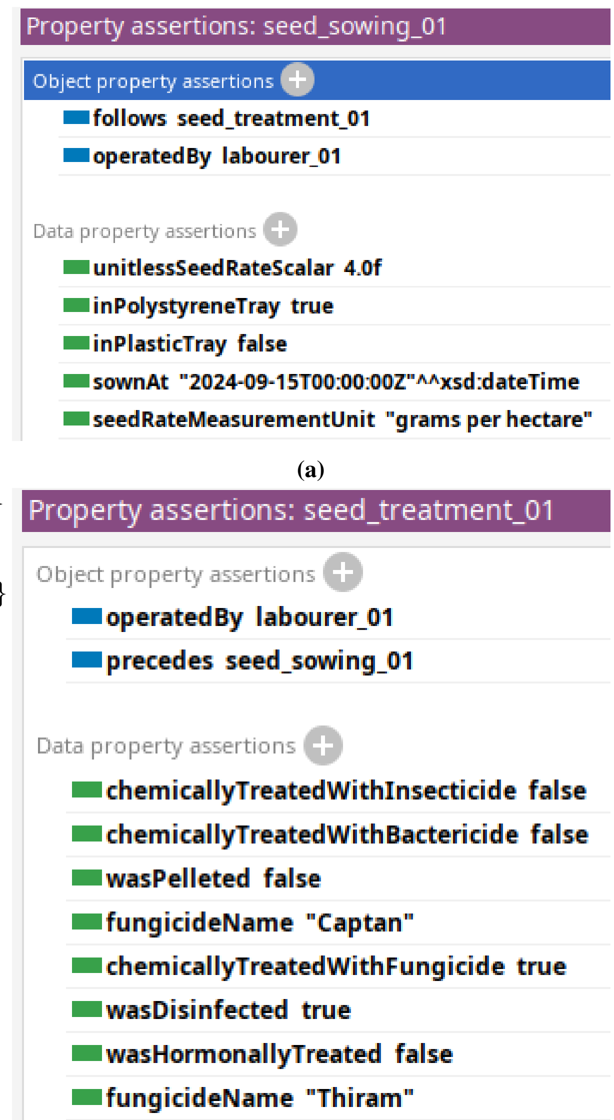
$$
\begin{aligned}
\text{TreatedSeed} \equiv \text{TobaccoSeed} \sqcap \\
& \big(\exists \text{chemicallyTreatedWithBactericide}.\{\text{true}\} \\
& \sqcup \exists \text{chemicallyTreatedWithFungicide}.\{\text{true}\} \\
& \sqcup \exists \text{chemicallyTreatedWithInsecticide}.\{\text{true}\} \\
& \sqcup \exists \text{wasDisinfected}.\{\text{true}\} \\
& \sqcup \exists \text{wasHormonallyTreated}.\{\text{true}\} \\
& \sqcup \exists \text{wasPelleted}.\{\text{true}\}\big)
\end{aligned}
$$

Figure 3 shows examples of individuals of the ***Sowing*** (a) and ***SeedTreatment*** (b) classes.

3.1.2 Practical Use Case Scenarios

To illustrate the practical utility of OnTop when integrated with a large language model, we present examples of real-life queries that farmers or agricultural advisors could submit:

- **Soil Management:** *"Given a field with clayey soil, high moisture, and pH 5.2, what lime application is recommended before planting tobacco?"*
  Response: Based on soil acidity and clay's moisture retention properties, the recommendation is a lime application of approximately 2.5 tons per hectare to raise pH to an optimal range between 6.0 and 6.5, improving nutrient availability and reducing aluminum toxicity.



(b)

Figure 3. Examples of individuals of sowing and seed treatment.

- **Pest and Disease Management:** *"In the topping stage, with humidity levels above 85%, what are the risks of fungal diseases, and what should be prioritized?"*
  Response: High humidity significantly increases the risk of blue mold (*Peronospora tabacina*). The recommendation includes immediate application of a fungicide appropriate for the growth stage, enhanced ventilation in the field, and continuous monitoring for early signs of infection.

- **Yield Forecast and Post-Harvest Management:** *"Given 120 mm of rainfall in the past 7 days combined with low sunlight, how might this affect curing and yield?"*
  Response: The excessive humidity may delay the curing process and increase the risk of mold formation. Recommendations include increasing ventilation during curing, avoiding overpacking leaves in curing barns, and adjusting harvest scheduling to minimize moisture-related losses.

These examples demonstrate the potential of OnTop to support decisions in different tobacco farming scenarios.

3.1.3 Metrics of OnTop

To evaluate OnTop, we used OntoMetrics [24][5]. OnTop presents considerable complexity and expressivity, with 1,291 axioms of which 880 are logical. The class count is 143, which provides an extensive framework for representing the various entities and relationships involved in tobacco production. OnTop's DL expressivity of SROIF(D) indicates that it supports role hierarchies, role inverses, role transitivity, and functionality, making it an effective tool for reasoning over complex relationships in the domain. This expressivity shows that OnTop can handle a suitable number of restrictions and can represent intricate class relationships and constraints.

OnTop has 96 object properties and 69 data properties, resulting in 165 properties. This shows that OnTop has a diverse representation of relationships (object properties) and attributes (data properties). OnTop contains 43 individuals, showing initial instances of the knowledge base. These individuals are concrete examples of the concepts and relationships defined in our ontology, providing a starting point for future knowledge expansion.

From a structural standpoint, the ontology contains 137 SubClassOf axioms and 11 EquivalentClass axioms, which reflect the detailed hierarchy and the definition of synonymous classes, respectively. The ontology also includes 17 transitive object properties and 13 inverse object properties that enhance its reasoning capabilities, enabling inferences across different levels of the class hierarchy. The presence of 101 object property domain axioms and 100 object property range axioms ensures that relationships are very detailed in the context of their respective domains and ranges. Moreover, the 16 proposed functional data property axioms guarantee that certain properties are constrained to have unique values, adding another layer of precision to the model.

The schema metrics suggest additional insights into the richness of our ontology. An inheritance richness of 0.958 and a relationship richness of 0.460 indicate OnTop captures a considerable quantity of hierarchical knowledge, with a high degree of inter-class relationships. The axiom-to-class ratio of 9.03 shows a moderately high level of detail in the axiomatization of each class, reflecting the ontology's depth in representing tobacco production knowledge.

3.2 Evaluation of OnTop Integration with a Large Language Model (LLM)

We consulted with an expert on agriculture field who confirmed that the questions posed are theoretically sound and relevant to the domain of tobacco production. Using these validated questions, we employed a Large Language Model (LLM) to test whether the structure of OnTop could enhance the capabilities of the undertesting LLM (UT-LLM) in providing accurate guidance. First, the two domain-specific questions were asked to the UT-LLM and the answers were retained. Following this, the OnTop OWL file was provided to the UT-LLM, and the same questions were asked again, this time with the ontology available. The responses were then analyzed to assess the extent to which the ontology influenced and improved the responses of the UT-LLM.

3.2.1 Application Scenario

We used ChatGPT-4o[6] as a UT-LLM to answer the following questions (Q):

> Q1) How does the treatment of seeds occur in tobacco production and what is its influence?
>
> Q2) How should tobacco harvesting work with regard to its phases and actions between these phases?

As the UT-LLM was verbose, we omitted part of the answer to make it easier to understand its reasoning. The main points of the answers (A) are the following (the highlights are ours):

> A1) [...] The seeds are extremely small, so **they are often coated to improve handling and sowing**. Key steps include: Seed Coating: Seeds are often coated with materials to make them easier to plant and manage. This helps in even distribution during sowing. Pre-germination Treatments: Seeds **may undergo pre-germination treatments**, including soaking in water or applying fungicides to prevent diseases. Controlled Environment for Germination: Tobacco seeds

[5] OntoMetrics is a web-based tool that validates and displays statistics about a given ontology. Link: https://ontometrics.informatik.uni-rostock.de/ontologymetrics/

[6] ChatGPT is an AI tool developed by the company OpenAI that provides chatbot functionality. GPT stands for Generative Pre-trained Transformer, a type of LLM.

are typically **germinated in nursery beds or greenhouses**, where temperature, humidity, and light are carefully controlled to ensure uniform growth.

A2) [...] Tobacco harvesting occurs in multiple phases depending on the type of tobacco and its intended use. The process generally follows these stages: Priming (Bottom Leaves): The first phase involves harvesting the bottom leaves, which are typically riper earlier. These leaves are crucial for certain types of tobacco products, like cigarettes. Farmers harvest them manually by pulling off the mature leaves. Mid-Stalk Leaves: Once the bottom leaves are harvested, attention turns to the middle stalk leaves, which are typically more valuable due to their higher nicotine content and better quality for curing. Top Leaves: The final phase is the harvesting of the top leaves, which mature last. These leaves are often thicker and contain more nicotine.

Actions Between Phases:

Curing: **After each phase, the leaves undergo curing** to preserve and enhance the flavor and aroma of the tobacco. Curing methods include air curing, flue curing, or sun curing, depending on the type of tobacco. Proper curing between harvest phases is vital for leaf quality. [...]

In the second part of the test, we sent to UT-LLM the OWL file of OnTop with the following message: "*Now, read integrally this ontology in the OWL file I am sending attached, and answer the same questions (Q1 and Q2)*". The main points of the enriched answers (Ae) are the following (the highlights are ours):

Ae1) [...] The treatment process includes **not only coating** but also protecting the seeds against potential threats such as insect and bacterial damage. [...] treatments may involve **insect control measures to ensure healthy seedlings**. Irrigation Systems: Once the seeds have germinated, **irrigation becomes a key factor**. [...] emphasizing the importance of maintaining proper water levels to support healthy seedling growth. Seedling Health Monitoring: Plant health is monitored throughout the early stages, ensuring that treatments are adjusted based on factors like **irrigation, pest control, and soil conditions**. Thus, seed treatment involves not only germination but also an ongoing system of health monitoring and environmental management, with a focus on protecting against pests and ensuring optimal growing conditions. This comprehensive approach influences both the growth rate and the quality of tobacco plants.

Ae2) [...] The first phase of harvesting is the picking of the bottom leaves, which mature earlier and are manually harvested. **Topping is another phase which refers to removing the flower from the plant**. This helps direct more energy to the leaves, enhancing their quality. [...]

Actions Between Phases:

Curing: Curing is a critical action between harvest phases, and additional focus must be applied to proper techniques to avoid issues **like mold formation on the leaves**. This reinforces the importance of careful environmental control during curing. [...] also important is a system of ongoing health monitoring during each phase, ensuring that any issues like **insect damage or under-curing** are addressed early.

3.2.2 Discussion on the Results

We consider the answers given by the UT-LLM before knowing OnTop (A1 and A2) to be reasonable. Regarding A1, it presents how seed treatments work and examples (e.g. soaking in water). It lacks detail, but it is overall trustworthy. Regarding (A2), it presents a simplified view of the process, which does not reflect the rigidity of the time frame for phases. When commenting on the actions between phases, the LLM gives an ambiguous and potentially incorrect answer, saying that *"after each phase, the leaves undergo curing"*.

After interacting with OnTop, in Ae1, the UT-LLM recognizes the importance of other treatments (e.g., insecticides) that were not mentioned in A1. It also goes tangentially off the question (which was asked only about seed treatment) and emphasizes the importance of irrigation, a key point in OnTop. Taking a more global approach than in the first answer, UT-LLM highlights the importance of irrigation, pest control, and soil conditions, so that seed treatments can have the desired effect.

In Ae2, it still simplifies the process but adds another layer of complexity, namely citing the topping process, which was absent in A2. About actions between phases, it still presents the potentially wrong statement *"curing is a critical action between harvest phases"*, which is negative (although here it especifies *"between **harvest** phases"*, which could be interpreted as: as leaves are harvested, they go to the cure process). In contrast, it recognizes the importance of avoiding mold (an important problem in tobacco curing, not mentioned in A2) and health monitoring to prevent insect damage and undercuring, both concepts present in OnTop.

To ensure the validity of our analysis, we again consulted with the agriculture expert, who reviewed our analysis of the answers provided by the UT-LLM both before and after interacting with OnTop. Her evaluation confirmed that the analysis of the LLM's responses was sound, with no significant theoretical errors identified. She also affirmed that improved responses after integrating OnTop added relevant agricultural insights, particularly with regard to irrigation, pest control, and disease management. Although acknowledging the minor mistake with respect to the curing process, she found that the LLM-enriched answers demonstrated a more comprehensive understanding of the tobacco production life cycle.

In summary, UT-LLM used OnTop knowledge to improve its responses, adding critical concepts, and providing a broader view of the whole tobacco lifecycle, even though it committed a mistake by suggesting curing between phases.

3.3 Conclusion

This paper presents the first formal ontology specifically designed for tobacco production, filling a critical gap in agricultural knowledge management for this economically significant crop. By modeling the entire tobacco production lifecycle, from seedbed treatment to post-harvest curing, the Ontology of Tobacco Production (OnTop) formalizes the complex relationships and processes of tobacco farming. OnTop was developed using OWL and integrates critical environmental and agronomic factors, allowing enhanced decision-making capabilities. OnTop could offer tobacco farmers, researchers, government officials, and advisors insight beyond traditional and experience-based practices, by incorporating real-time and historical data from geographic information systems and precision agriculture tools.

Although OnTop presents a robust formalization of the tobacco production lifecycle, some limitations persist. Firstly, its current scope primarily focuses on the production stages, with limited modeling of post-harvest logistics, processing, and supply chain dynamics. Secondly, the knowledge embedded in OnTop represents a snapshot in time; agricultural practices, pest dynamics, climate patterns, and regulatory requirements evolve, which imposes a challenge for the continuous updating and maintenance of the ontology.

Another limitation concerns the adoption of ontology-based systems in farming communities. Many farmers and advisors lack familiarity with semantic technologies. For widespread adoption, user-friendly interfaces that abstract away technical complexity are essential. Furthermore, the practical deployment of OnTop requires integration with local data sources, such as soil databases, weather stations, and precision agriculture sensors, which may not be readily available in all regions.

Lastly, while the evaluation using an LLM demonstrates promising improvements in the quality of recommendations, further testing in operational, real-world contexts is necessary to ensure reliability, trustworthiness, and relevance for diverse farming environments.

Addressing these challenges will require technical refinement, stakeholder engagement strategies, capacity-building initiatives, and possibly creating governance models for ontology updates.

A critical next step involves the empirical validation of OnTop in real-world tobacco farming environments. This validation may be conducted through partnerships with tobacco cooperatives, agricultural consultants, and farmers. The process would involve deploying OnTop integrated into a decision support system, followed by structured interactions with end-users, including farmers, agronomists, and technical advisors.

The roadmap for validation includes the following stages: i) *Pilot Studies:* Perform pilot deployments with small groups of farmers to evaluate the usability and decision support effectiveness; ii) *User-Centered Evaluation:* Collect feedback from farmers and agronomists regarding the accuracy, usefulness, and interpretability of the recommendations; iii) *Iterative Refinement:* Based on feedback, refine the ontology structure, expand the coverage of classes, and adjust relationships or data properties if necessary; iv) *Quantitative Evaluation:* Measure the impact of the ontology-based system on key agronomic metrics, such as yield, input efficiency, and pest/disease management effectiveness.

Future work will focus on expanding OnTop to cover more granular details of post-harvest processes and incorporating additional environment variables (e.g., microclimate and soil microbiome data) to enhance the decision-making capabilities of the system. In addition, a more rigorous testing phase will involve applying OnTop in other real farming scenarios to validate its effectiveness in real-time decision support. Integrating it into more advanced machine learning models and AI systems (beyond LLMs) could enhance the predictive power of our ontology, making it a more valuable tool for precision agriculture. Large databases of the tobacco industry would provide a rich testing environment for the ontology. For example, historical and real-time climate data could be integrated to enrich the decision process.

Conflict of Interest Statement

The authors declare that they have no conflicts of interest related to this research.

Acknowledgements

The authors would like to express their sincere gratitude to Dr. Herica Fernanda de Sousa Carvalho, Ph.D. in the field of Agriculture from the Federal University of Pernambuco (UFPE) and researcher of the Renato Archer Information Technology Center (CTI), for her contribution in reviewing the domain-specific questions used in the LLM test. Her expertise in agricultural sciences provided essential insights and helped ensure the theoretical soundness of the evaluation process.

Author contributions

Luis Felipe Medeiro Alves: Conceptualization; Methodology; Investigation; Ontology Development; Formal Analysis; LLM enrichment and test; Writing and Editing the Original Draft; Project Administration. **José Maria Parente de Oliveira:** Investigation; Ontology Review and Validation; Formal Analysis; LLM test review; Reviewing and Editing the Manuscript. **Rodrigo Bonacin:** Investigation; Ontology Review and Validation; Formal Analysis; LLM test review; Reviewing and Editing the Manuscript. **Ferrucio de Franco Rosa:** Conceptualization; Methodology; Investigation; Ontology Review and Validation; Formal Analysis; LLM test review; Writing and Editing the Original Draft; Research Supervision.

References

[1] FAOSTAT. *Food and Agriculture Organization of the United Nations*. 2020. ⟨http://www.fao.org/faostat/en/⟩. Retrieved May 17, 2020.

[2] Frente Parlamentar da Agropecuária. *Cultura do tabaco*. 2020. ⟨https://fpagropecuaria.org.br/2020/06/29/cultura-do-tabaco/?pdf=30788⟩.

[3] Confederação da Agricultura e Pecuária do Brasil. *Produção de fumo cresce 579% em um ano no agreste*. 2020. ⟨https://www.cnabrasil.org.br/noticias/producao-de-fumo-cresce-579-em-um-ano-no-agreste⟩.

[4] GEIST, H. J. et al. Tobacco growers at the crossroads: Towards a comparison of diversification and ecosystem impacts. *Land Use Policy*, v. 26, n. 4, p. 1066–1079, 2009.

[5] ALVES, L. F. M.; FRANCO, F. d. R. Gis-based recommender systems for agriculture: A systematic literature review. Zenodo. https://doi.org/10.5281/zenodo.16579831. 2025.

[6] KITCHENHAM, B. *Procedures for performing systematic reviews*. [S.l.], 2004. 28 p.

[7] PIERCE, F. J.; CLAY, D. E.; SCHULER, R. T. *Applications of GIS in precision agriculture*. [S.l.]: John Wiley and Sons, 2007.

[8] ANLEY, M. B.; TESEMA, T. B. A collaborative approach to build a kbs for crop selection: Combining experts knowledge and machine learning knowledge discovery. In: *Communications in Computer and Information Science*. [S.l.]: Springer International Publishing, 2019. p. 80–92.

[9] DABRE, K. R.; LOPES, H. R.; DMONTE, S. S. Intelligent decision support system for smart agriculture. In: *2018 International Conference on Smart City and Emerging Technology (ICSCET)*. [S.l.]: IEEE, 2018.

[10] FARHEEN, N. Rainfall prediction and suitable crop suggestion using machine learning prediction algorithms. In: *Advances in Intelligent Systems and Computing*. [S.l.]: Springer Nature Singapore, 2021. p. 497–513.

[11] FEGADE, T. K.; PAWAR, B. V. Crop prediction using artificial neural network and support vector machine. In: *Data Management, Analytics and Innovation*. [S.l.]: Springer Singapore, 2020. p. 311–324.

[12] RAHMAN, S. A. Z.; MITRA, K. C.; ISLAM, S. M. M. Soil classification using machine learning methods and crop suggestion based on soil series. In: *2018 21st International Conference of Computer and Information Technology (ICCIT)*. [S.l.]: IEEE, 2018.

[13] AVEZBOYEV, S.; SHARIPOV, S.; XUJAKELDIEV, K. Development of projects for recultivation of lands using gis technologies. In: *IOP Conference Series: Earth and Environmental Science*. [S.l.: s.n.], 2023. v. 1138, n. 1, p. 012019.

[14] GAO, Z.; LI, Z.; CHEN, J. Planning and suggestions for the walnut production areas in beijing. In: *2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. [S.l.]: IEEE, 2018.

[15] RAJAMANI, K.; HARI, N.; RAJASHEKAR, M. Soil fertility evaluation and gps-gis-based soil nutrient mapping of krishi vigyan kendra, palem, telangana. *International Research Journal of Pure and Applied Chemistry*, p. 139–145, 2020.

[16] RAMCHANDANI, L.; PATEL, S.; SUTHAR, K. A multiple criteria-based context-aware recommendation system for agro-cloud. In: *Advances in Computing and Intelligent Systems*. [S.l.]: Springer Singapore, 2020. p. 557–564.

[17] RODRÍGUEZ-GARCÍA, M. ; GARCÍA-SÁNCHEZ, F. Croppesto: An ontology model for identifying and managing plant pests and diseases. In: *Communications in Computer and Information Science*. [S.l.]: Springer International Publishing, 2020. p. 18–29.

[18] SUCHITHRA, M. S.; PAI, M. L. Data mining-based geospatial clustering for suitable recommendation system. In: *2020 International Conference on Inventive Computation Technologies (ICICT)*. [S.l.]: IEEE, 2020.

[19] MAWARDI, S. Y. et al. An ontology-driven decision support system for wheat production. *International Journal of Computer Science and Telecommunications*, v. 4, n. 8, p. 11, 2013.

[20] COX, S. et al. Toward an ontology of tobacco, nicotine and vaping products. *Addiction*, v. 118, n. 1, p. 177–188, 2023.

[21] NOTLEY, C. et al. Toward an ontology of identity-related constructs in addiction, with examples from nicotine and tobacco research. *Addiction*, v. 118, n. 3, 2022.

[22] International Organization for Standardization. *Information technology — Top-level ontologies (TLO) — Part 2: Basic Formal Ontology (BFO)*. [S.l.], 2016. ISO Standard No. 21838-2:2020.

[23] ALVES, L. F. M.; ROSA, F. F. *OnTop ontology*. 2024. ⟨https://github.com/lfmalves/OnTop⟩.

[24] LANTOW, B. Ontometrics: Putting metrics into use for ontology evaluation. In: *International Conference on Knowledge Engineering and Ontology Development*. [S.l.: s.n.], 2016.