CarboFarm: Data Integration and Knowledge Generation for Agricultural GHG Inventories

Luiz Santos[1], Regina Braga[1], José Maria N. David[1]
[1]Department of Computer Science - Federal University of Juiz de Fora (UFJF).
Postal Code 20.010 – 36036-900 – Juiz de Fora – MG – Brazil

fernando.santos@estudante.ufjf.br, regina.braga@ufjf.br (advisor), jose.david@ufjf.br(co-advisor)

Abstract. *This work presents an architecture called CarboFarm, that encompasses an ontology for the syntactic and semantic integration of heterogeneous databases related to the agricultural domain. By utilizing ontology, we contribute to the standardization and interpretation of domain concepts and add semantic information to support the generation of GHG (Greenhouse Gas) inventories on farms. The Artificial Intelligence (AI) component processes integrated data to derive GHG information. The inventories can identify imbalances between emissions and gas stocks, helping search for solutions to neutralize emissions. We conducted a case study, showing that CarboFarm can support the balancing of gases and the generation of carbon credits.*

Resumo. *Este trabalho apresenta uma arquitetura, chamada CarboFarm, que engloba uma ontologia, para integração sintática e semântica de bancos de dados heterogêneos relacionados a dados agrícolas. Com a ontologia, pretendemos contribuir para a padronização e interpretação de conceitos do domínio, adicionando informações semânticas para suporte à geração de inventários de GEE (Gases de Efeito Estufa) em fazendas. Um componente de Inteligência Artificial (IA) processa os dados integrados para derivar informações de GEE. Os inventários podem identificar desequilíbrios entre emissões e estoques de gases, auxiliando na busca de soluções para neutralizar as emissões. Realizamos um estudo de caso, mostrando que CarboFarm pode auxiliar no balanceamento dos gases e na geração de créditos de carbono.*

1. Dissertation Defense Data and Highlights

Date of Defense and Approval: 23/09/2024.
Posgraduate Program in Computer Science (PPCC) - Federal University of Juiz de Fora (UFJF)/Brazil.
Category: Master's degree.
Author: Luiz Fernando Santos.
Advisors: - Regina Braga (regina.braga@ufjf.br) – advisor;
- José Maria David (jose.david@ufjf.br) – co-advisor.
Board Members: - Victor Stroele (victor.stroele@ufjf.br), UFJF (internal member);
- Marta Mattoso (marta@cos.ufrj.br) – UFRJ (external member).
Note and honorable mention: UFJF assigns the grade "Approved", and does not note or mention praise or anything equivalent.

Highlights:
- Ontology to syntactically and semantically integrate GHG agricultural inventory data.
- Machine learning techniques to generate knowledge from ontology-integrated data.
- Machine learning predictions to contribute to decision-making on farms.
- Analysis and extraction of geospatial dataset to produce GHG inventory on farms.
- Cloud architecture for supporting GHG inventories and generating carbon credits.

2. Characterization of Research Problem and Motivation

Global warming and climate change have been topics of great interest in recent years, as it is related to greenhouse gas (GHG) emissions. The agricultural sector suffers the consequences of these changes. However, it is also one of the top global emitters of GHG. Due to the importance of agricultural activities for food systems, this sector is a fundamental part of the GHG emission mitigation strategy. This is a complex sector in its environmental, social, and economic aspects. There is a need to propose new solutions that provide more sustainable production. In the farm environment, an important step is the generation of GHG inventories, based on heterogenous data that came from different sources. Based on the knowledge generated by this integrated data, problems can be identified, and solutions can be searched to increase carbon sequestration and reduce emissions. A positive carbon balance generates carbon credits with economic return.

Public datasets and datasets collected on rural properties, when available, can contribute to the generation of inventories and the promotion of more sustainable agricultural practices. This study presents an architectural proposal called CarboFarm. CarboFarm can help small-scale farmers to quantify carbon emissions and analyze crop alternatives suitable for their region. The aim is to trade carbon credits when the market is regulated in Brazil. Considering COP30 in Brazil, solutions like this meet Brazil's desire to be an important agent of sustainability in agricultural practices.

The Main Research Question (RQ) addressed in this work is: "*How does integrating data from GHG emissions and stocks support the generation of agricultural inventories?*". To help to answer the RQ, Secondary Research Questions (SRQ) were proposed: SRQ1: "*How does integrating data from emissions and GHG stock sources support more sustainable farm production?*"; SRQ2: "*Can knowledge be extracted for generating carbon credits from integrating data on emission sources and GHG stocks?"*

3. Objectives

This work presents an architecture called CarboFarm for integration of agricultural data. The objective is to generate greenhouse gas inventories on farms. The integrated data promote the generation of knowledge to support rural landowners' decision-making and the generation of carbon credits. CarboFarm allows data integration from heterogeneous sources, including datasets, deforestation monitoring alerts, and sensor data. CarboFarm aims to provide information or be integrated into MRV systems and decision support applications for rural landowners.

This dissertation´s main objective is to explore the data extraction, integration, and analysis services. An ontological model allows syntactic and semantic integration, contributing to data standardization, sharing, and interoperability. The analysis of historical data through machine learning techniques allows efficient processing of large volumes of data. It aims to identify patterns, trends, and insights that can be useful for decision-making. This provides knowledge to choose the best conditions of soil use and cultivation techniques. To support our approach, we carried out a case study integrating datasets of GHG emissions and stocks on Brazilian rural properties. To achieve these results, the following objectives were considered: i) syntactically and semantically integrate heterogeneous data sets related to emission sources and GHG stocks on farms; ii) using integrated data to support the generation of agricultural GHG inventories; iii) using integrated data to generate knowledge to support decision-making and generate carbon credits.

4. Contributions

As a specific contribution, this research proposes: i) An ontological model for data integration focusing on the standardization and interoperability of information between measurement systems; ii) An AI component designed to process integrated data, derive information and generate knowledge; iii) Using the knowledge generated to support constructing GHG inventories, our solution can facilitate decision-making, and assist the generation of carbon credits. Throughout the research, we did not find studies that addressed ontologies and AI components built to address climate issues with an emphasis on GHG inventories in agriculture. CarboFarm can help to fill this gap, offering integrated data for AI components that aim to generate knowledge to support rural producers' decisions in constructing GHG inventories and generating carbon credits.

Considering COP30, solutions like this meet Brazil's desire to be an important agent of sustainability in agricultural practices, with the help of intelligent data processing.

5. State-of-the-art advances

Our work involves technical, environmental, economic and social perspectives. In our research, we consider some state-of-the-art advances, listed below.

4.1) Using ontology in this domain contributes to generating more detailed GHG inventories. The addition of semantic information contributes to the generation of knowledge, which helps rural landowners in making decisions. Furthermore, the ontological model can make a decisive contribution to interoperability between MRV systems through standardization and interpretation of the meaning of terms, eliminating or reducing conceptual and terminological confusion. Standardizing of terms would be important for creating applications for the carbon market, offering greater reliability and transparency. From a database research perspective, a well-defined semantic structure could provide better system specifications and data reuse.

4.2) Using machine learning techniques, we were able to generate regression models that indicate the best use of land for a rural property, considering the type of cultivation desired, location (municipality), and climate.

4.3) The CarboFarm architecture is intended to integrate data in a cloud environment. For the context of applications in the agricultural domain, it is necessary to consider the target audience, mainly small farmers, who may have access difficulties, whether of a technical nature, such as handling technologies (computers, software, use of the Internet), cognitive order (autonomy and independence in the use of technologies) or economic order (ability to acquire more powerful computing equipment and have every time Internet connectivity). Cloud architecture also facilitates reuse by other countries that adopt the same inventory generation methodology and have agricultural GHG datasets. Some South American and Southeast Asian countries could use Brazilian estimates, as they have climate characteristics similar to Brazil's [MapBiomas 2021].

6. Evaluation

We conducted a case study with integrated public datasets from institutions in Brazil and detailed the land cover and use balance of Brazilian farms. Data obtained from rural properties can contribute to generating GHG inventories. But it must be integrated. This integrated data not only fosters a deeper understanding of the best practices in agriculture, making them more sustainable, but also contributes to the creation of carbon credit. Land use and cover for Brazilian rural properties were obtained through the integration of rural property polygons (shapefiles) obtained from the Land Tenure Map [De Freitas *et al.* 2018] and land use and cover maps of the MapBiomas Project [MapBiomas 2021]. Other integrated data source was the soil carbon stock from the entire Brazilian territory [MapBiomas 2023]. Data on carbon emission estimates from land use and cover were extracted from the BRLUC (*Brazilian Use Change*) Method [BRLUC 2022]. Data on Brazilian cities and biomes were obtained from the website of the Brazilian Institute of Geography and Statistics (IBGE). Our case study sampled 91,747 properties corresponding to 295,854 cultivation areas. Once GHG emission estimates have been calculated in each area, they are consolidated to generate estimates for the entire property.

7. Produção Científica

7.1) Artigo: "Uma abordagem para suporte à decisão no processo de geração de créditos de carbono em propriedades rurais".

- Artigo publicado no Simpósio Brasileiro de Sistemas Colaborativos (SBSC2023)
- DOI: https://doi.org/10.5753/sbsc.2023.229059
- Este foi o primeiro trabalho publicado, apresentando os conceitos inicialmente trabalhados, mostrando resultados iniciais e a viabilidade de continuação da pesquisa.

7.2) Artigo: "Towards a SECO for Carbon Credit Control".

- Artigo publicado no "*11th ACM/IEEE International Workshop on Software Engineering for Systems-of-Systems and Software Ecosystems - SESoS@ICSE 2023".*
- DOI: https://doi.org/10.1109/SESoS59159.2023.00008
- 182 visualizações de texto completo (julho de 2023 a junho de 2025)
- Este trabalho teve o objetivo de apresentar uma proposta de *framework* para o desenvolvimento de aplicações relacionadas ao controle de emissões de Gases de Efeito Estufa (GEE) e a geração de créditos de carbono no domínio da agropecuária.

7.3) Artigo: "CarbOnto: Data Integration Toward Net Zero".

- Artigo publicado na *IEEE Access, vol. 12, pp. 148783-148795, 2024.*
- DOI: https://doi.org/10.1109/ACCESS.2024.3477259
- 362 visualizações de texto completo (outubro de 2024 a junho de 2025), Web of Science: 15, Scopus: 15.
- Este trabalho teve o objetivo de apresentar a ontologia CarbOnto, um componente da arquitetura CarboFarm desenvolvida na pesquisa da dissertação. CarbOnto é uma ontologia destinada a integração sintática e semântica de fontes de dados heterogêneas, relacionadas a dados de emissão e estoques de GEE em fazendas.

7.4) Menção Honrosa: Dissertação ganhou menção honrosa no CTD do SBSI 2025.

7.5) Softwares produzidos

- CarbOnto - ontologia para integração de dados de bases heterogêneas de GEE agrícolas. Disponível em *https://github.com/LFS19/CarbOnto.*
- CarboFarm - arquitetura em camadas com a finalidade de: (i) integração de dados por meio da ontologia CarbOnto; (ii) análise por meio de algoritmos de aprendizagem de máquina; (iii) oferecimento de suporte a decisão para produtores rurais. Disponível em *https://github.com/LFS19/Carbofarm.*

Referências

Brazilian Land Use Change (BRLUC) (2022). Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA). Available: https://brluc.cnpma.embrapa.br.

De Freitas, F. L. M.; Guidotti, V.; Sparovek, G.; hamamura, C. (2018). "Land Tenure Map of Brazil. Atlas - A Geografia da Agropecuária Brasileira". IMAFLORA: Piracicaba, Brazil, 1812, 5, 2018. Available: https://bit.ly/tmapbrazil.

Mapbiomas (2021). "Collection of the Annual Series of Land Use and Cover Maps of Brazil." Available: https://plataforma.brasil.mapbiomas.org.

MapBiomas. (2023). "Mapeamento anual do estoque de carbono orgânico do solo no Brasil 1985-2021 (coleção beta)", "Annual mapping of soil organic carbon stock in Brazil 1985-2021 (beta collection)", DOI: 10.58053/MapBiomas/DHAYLZ, MapBiomas Data, V1".