ISSN: 2447-3359

REVISTA DE GEOCIÊNCIAS DO NORDESTE

Northeast Geosciences Journal

v. 11, nº 2 (2025)

https://doi.org/10.21680/2447-3359.2025v11n2ID39805

Enhancing interoperability between geospatial data: NPL semantic similarity alignment metrics with AI between land cover and land use data

Aprimorando a interoperabilidade entre dados geoespaciais: métricas de alinhamento de similaridade semântica NPL com IA entre dados de cobertura e uso da terra

Vitor Silva de Araujo [1]; Silvana Phillipi Camboim [2]; Naíssa Batista da Luz [3]

[1] Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas/Departamento de Geomática, Curitiba/PR, Brasil. Email: vitorsilvadearaujo@ufpr.br
ORCID: http://orcid.org/0000-0003-4880-3016

[2] Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas/Departamento de Geomática, Curitiba/PR, Brasil. Email: silvanacamboim@ufpr.br
ORCID: https://orcid.org/0000-0003-3557-5341

[3] Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas/Departamento de Geomática, Curitiba/PR, Brasil. Email: naissa@ufpr.br
ORCID: https://orcid.org/0000-0001-9803-9170

Abstract: The evolution of geospatial data sources and their diverse classification systems poses challenges to data integration and interoperability. This research addresses these challenges by introducing an AI-driven methodology that utilises Natural Language Processing (NLP) to measure semantic similarity between land use, vegetation classification systems, and the national topographic database. Leveraging NLP techniques, such as those in ChatGPT-4.0, this approach automates the semantic alignment process, reducing manual work. The study aimed to align the Brazilian ET-EDGV topographic mapping with broader national (IBGE Vegetation and Land Use Manuals) and international (Dynamic World, Global Forest Resources Assessments (FRA)) classification systems. By applying semantic similarity coefficients, the research sought to create a harmonized framework for integrating geospatial data. The methodology combined AI-based semantic similarity measures, ensuring consistent data alignment. Results showed strong alignments for classes like “Cultivated Vegetation” and “Crops” and identified challenges for unique Brazilian ecosystems such as “Campinarana”. The “Mangrove” class highlighted the need for context-specific definitions. The study concludes that NLP can contribute to automated semantic alignment, enhancing geospatial data integration and interoperability. Although focused on Brazilian data, this methodology is adaptable globally, supporting more accurate landscape representation and informed decision-making. Future research should integrate advanced AI models and broader ecosystems to refine the process.

Keywords: Topographic Map; Natural language processing.; Semantic similarity.

Resumo: A evolução das fontes de dados geoespaciais e seus variados sistemas de classificação apresentam desafios de integração e interoperabilidade de dados. Esta pesquisa aborda esses desafios introduzindo uma metodologia orientada por IA usando Processamento de Linguagem Natural (PLN) para medir a similaridade semântica entre o uso da terra, sistemas de classificação de vegetação e bancos de dados topográficos nacionais. Aproveitando técnicas de PNL, como as do ChatGPT-4.0, esta abordagem automatiza o processo de alinhamento semântico, reduzindo o trabalho manual. O estudo teve como objetivo alinhar o mapeamento topográfico brasileiro ET-EDGV com sistemas de classificação nacionais mais amplos (Manuais de Vegetação e Uso da Terra do IBGE) e internacionais (Dynamic World, Global Forest Resources Assessments (FRA)). Ao aplicar coeficientes de similaridade semântica (valores S), a pesquisa buscou criar uma estrutura harmonizada para integrar dados geoespaciais. A metodologia combinou medidas de similaridade semântica baseadas em IA, garantindo alinhamento consistente de dados. Os resultados mostraram fortes alinhamentos para classes como "Vegetação Cultivada" e "Culturas" e identificaram desafios para ecossistemas brasileiros únicos, como "Campinarana". A classe “Mangrove” destacou a necessidade de definições específicas de contexto. O estudo conclui que o NLP pode contribuir para o alinhamento semântico automatizado, aprimorando a integração e a interoperabilidade de dados geoespaciais. Embora focada em dados brasileiros, essa metodologia é adaptável globalmente, apoiando melhor representação da paisagem e tomada de decisão. Pesquisas futuras devem integrar modelos avançados de IA e ecossistemas mais amplos para refinar o processo.

Palavras-chave: Mapa topográfico; Processamento de linguagem natural; Similaridade semântica.

Received: 09/04/2025; Accepted: 03/10/2025; Published: 25/12/2025.

1. Introduction

Describing the landscape is a fundamental function of mapping, with topographic mapping specifically addressing the representation of the landscape. Fremlin and Robinson (1998) state that topographic mapping represents the Earth as a composite entity, where the landscape reflects its appearance. However, major landscape elements, such as vegetation, have consistently posed challenges due to the tendency for rapid obsolescence (Gersmehl, 1981; Langran, 1985). The advent of remote sensing technologies early on demonstrated their value for landscape mapping (Doyle, 1973). Initially, LULC classification adhered to two significant principles: scale-based hierarchization and semantic compatibility with other authoritative data sources (Andeson et al., 1976).

Over time, many semantic definitions have emerged from various national systematic mapping initiatives and numerous land use and land cover (LULC) monitoring projects. Today's landscape demands data integration from multiple sources, making semantic compatibility essential for achieving full interoperability. Integrating semantics into mapping—whether for LULC or topographic purposes—requires careful alignment with the classifications employed by various regulations. Integrating heterogeneous sources requires semantic compatibility, especially for vague concepts like "forest" (Bennett, 2001, 2001; Mallenby, 2008; Varzi, 2001).

Integrating and reusing data from many sources, scales, and uses depends on data interoperability. Ballatore et al. (2013) and Robinson et al. (2017) emphasize that cartographic items' clarity, dependability, and applicability can be negatively impacted if models lack conceptual alignment. Digital, interactive, and user-oriented contexts especially aggravate this problem. Another problem is the rare updates to Brazilian topographic maps; many still rely on data from the 1990s. This contradicts the quick regional changes in biomes like the Amazon, Cerrado, and Caatinga (Souza, 2020), and emphasizes the need for integrated, flexible responses. Research reveals that geospatial data integration calls for structural and geometric compatibility. More importantly, it also calls for conceptual coherence between data models (Kuhn, 2003; Yu et al., 2018; Machado and Camboim, 2024). This is particularly crucial when dealing with complex thematic categories, such as vegetation, terrain, or land cover.

In Brazil, there is a clear lack of methodologies for automated data integration between institutions, such as the Brazilian Army Geographical Service Department (DSG) and Brazilian Institute of Geography and Statistics (IBGE), which map similar or equivalent concepts. As Souza et al. (2025) highlight, improvements in textual analysis, natural language processing (NLP), and artificial intelligence technologies could help address some issues with semantic alignment between various conceptual frameworks. The definitions of these categories have long differed across institutions, regions, and technical disciplines (Bravo, 2014; Brown et al., 2022). With subjective decisions about concepts, this alignment is now a manual, time-consuming process. Currently, there are no specialized tools to measure or quantify the equivalence between entities mapped in different models.

Quantifying equivalence, expressed as semantic similarity, between entities mapped to different data models is expected to allow for class alignment based on their values. Thus, similar semantic definitions should have metrics that indicate a high degree of equivalence between two entities. Similarly, it will be possible to analyze overall metrics across models, quantifying equivalence between them and highlighting the least and most equivalent entities based on individual analysis, which can indicate which data characteristics can foster interoperability when adapted. Furthermore, it indicates a path to automating this alignment process.

This not only addresses the gap in national topographic map coverage by suggesting the use of data from more frequent mappings and with metric similarity, but also opens up the possibility of improving the current model, using similar or different metric indicators between entities. This opens the possibility of deepening interoperability between institutions, where, in a beneficial scenario, these national institutions would have a greater degree of interoperability in their productions, fostering evidence-based public policies.

In this context, Natural Language Processing (NLP) emerges as a critical tool to formalize geosemantics Kuhn (2005), allowing for the comparison of definitions and improving interoperability, which is critical for integrating varied geospatial datasets (Elavarasi et al., 2014; Martinez-Gil, 2022; Meng et al., 2013; Yuhua Li et al., 2003). As new data sources become available, NLP techniques offer a powerful approach for processing, understanding, and reconciling the extensive text-based components of geospatial information. This study utilises NLP and AI to align the Brazilian topographic database

(ET-EDGV) with national and international classifications, thereby combining more frequently updated data and enhancing the interoperability and quality of geospatial information.

The primary objective is to achieve semantic alignment through similarity between the Brazilian topographic database and national and international databases, with the potential to automate the operation. The methodology is illustrated by a Brazilian case study, whose topographic data model is characterized by high positional accuracy but is not sufficiently updated (Silva & Camboim, 2020). In contrast, other LULC databases, updated more frequently using satellite imagery, provide more current data. Integrating these datasets at a semantic level would provide significant benefits for conservation and decision-making by combining the strengths of both data types.

The results suggest the possibility of semantically aligning different data models, metricizing operations by similarity, and enabling future process automation. Therefore, in addition to the alignments, it was possible to analyze the class groupings and their differences, indicating the possibility of adapting the data to promote interoperability.

This research advances data integration practices and enhances the quality of geographic information. Although focused on Brazilian geospatial data, the proposed methodology has broad applicability. It provides a robust framework for measuring semantic similarity, which can inform the integration of diverse geospatial data sources globally, thereby fostering data interoperability and enhancing the overall quality of geospatial information.

2. Methodology

2.1 Selection and Application of Semantic Similarity Methods

To measure semantic similarity in geospatial data, methods have been developed: knowledge-based methods, corpus-based methods, deep neural network-based methods, and hybrid methods (Gorman & Curran, 2006; Rada et al., 1989; Sánchez et al., 2012; Wang et al., 2017; Zhu & Iglesias, 2017). Among these, corpus-based methods were selected for their ability to leverage large volumes of text and capture contextual relationships between geospatial terms, based on the distributional hypothesis (Ali et al., 2018; Chandrasekaran & Mago, 2022; Martinez-Gil, 2022; Sitikhu et al., 2019; Gorman & Curran, 2006). Using this hypothesis, corpus-based methods construct vector representations that effectively capture semantic relationships within geospatial terminology.

Word embeddings have gained prominence among the available corpus-based techniques. Various methods such as neural networks and word co-occurrence matrices have been used to generate these embeddings, with popular models including word2vec, GloVe, fastText and BERT (Bojanowski et al., 2017; Devlin et al., 2019; Levy & Goldberg, 2014; Mikolov et al., 2013; Pennington et al., 2014; Schnabel et al., 2015). Evaluating the effectiveness of these models (Dharma et al., 2022), we selected the Generative Pre-trained Transformer (GPT) model, the latest advancement in NLP, which is known for its ability to capture complex semantic nuances in large datasets. This choice reflects the effectiveness and robustness of the Transformer architecture (Vaswani et al., 2023), which handles semantic relationships in the geospatial domain.

Once the word vectors are generated, the next critical step is accurately measuring their distance. Among various similarity measures, cosine similarity is the most effective for complex geospatial concepts (Ali et al., 2018; Chandrasekaran & Mago, 2022; Machado-García et al., 2014; Sitikhu et al., 2019). It is widely used in NLP due to its ability to compare vectors of different lengths and capture directional relationships, making it ideal for evaluating semantic similarities in geospatial data.

Cosine similarity calculates the cosine of the angle between two vectors, with values ranging from -1 (opposite vectors) to 1 (identical vectors). This measure involves normalizing vectors and calculating their dot product, providing a reliable indication of textual relationships in vector space (Manning et al., 2009; Wilson & Schakel, 2015). This equation was originally used in an information retrieval context and is now adapted for comparing semantic definitions

Equation 1 shows the formula for cosine similarity:

$$sim(d_1, d_2) = \frac{\overrightarrow{V}(d1).\ \overrightarrow{V}(d2)}{\left|\overrightarrow{V}(d1).\ |x|\ \overrightarrow{V}(d2)\right|}$$

Equation 1 – Cosine similarity equation.
Source: Manning et al. (2009).

$\overrightarrow{V}$(d1) and $\overrightarrow{V}$(d2) represent the vector representations of documents d1 and d2, and $\left|\overrightarrow{V}(d1)\right|$ and $\left|\overrightarrow{V}(d2)\right|$ are their Euclidean lengths. This normalization ensures that cosine similarity focuses on the direction of the vectors, disregarding their absolute magnitude, thus making it particularly effective for comparing definitions from different geospatial data sources. Figure 1 illustrates the components used to determine the similarity between $\overrightarrow{V}$(d1) and $\overrightarrow{V}$(d2), where $\overrightarrow{V}(q)$ represents the query vector, and θ is the angle between them.

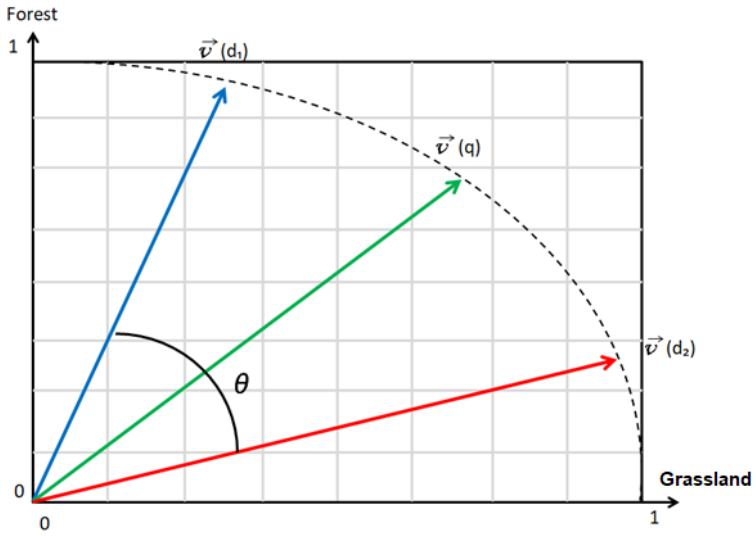


Figure 1 - Cosine similarity components illustrated between d_1 and d_2. sim(d1, d2) = cos θ.
Source: Adapted from MANNING et al., 2009.

To apply this method, a query is treated as a "word box", and cosine similarity is used to measure the score of a definition against that query. This approach allows the selection of top-scoring matches based on their similarity (Manning et al., 2009). Equation 2 shows how the score of cosine similarity for a given query and document is computed:

$$Score\,(q, d) = \frac{\overrightarrow{V}(q).\ \overrightarrow{V}(d)}{\left|\overrightarrow{V}(q).\ \overrightarrow{V}(d)\right|}$$

Equation 2 – Score of the cosine similarity equation.
Source: MANNING et al. (2009).

The selection of corpus-based semantic similarity methods, specifically employing GPT-generated embeddings and cosine similarity, has been examined and validated as an appropriate approach for measuring semantic similarity in

geospatial data. This combination provides a robust solution that can effectively handle the complexities inherent in geospatial terminology, thereby significantly enhancing data interoperability and the overall quality of geographic information.

2.2 Case Study Definition and Data Collections

In this case study, the methodology was applied to enhance the national topographic database at a 1:25,000 scale, which serves general-purpose mapping needs differently than thematic maps designed for specific uses (Anderson et al., 1976; Doyle, 1973; Fremlin & Robinson, 1998). Consequently, this section justifies the selection of classes from other models, which are used as inputs for the method, by prioritizing the legend of the topographic mapping. Additionally, a proposal for harmonizing concept definitions from various sources is presented, aiming to establish more accurate semantic definitions for LULC in the topographic map classes and to reduce ambiguity and vagueness of these geographical concepts.

This methodology was developed to semantically align the definitions of the LULC classification with those of the Brazilian topographic mapping. The process involves three main steps: selecting compatible data sources, collecting semantic data, and conducting data processing and alignment. Each step is detailed in the methodological flowchart in Figure 2, which outlines the computational routine and input data used.

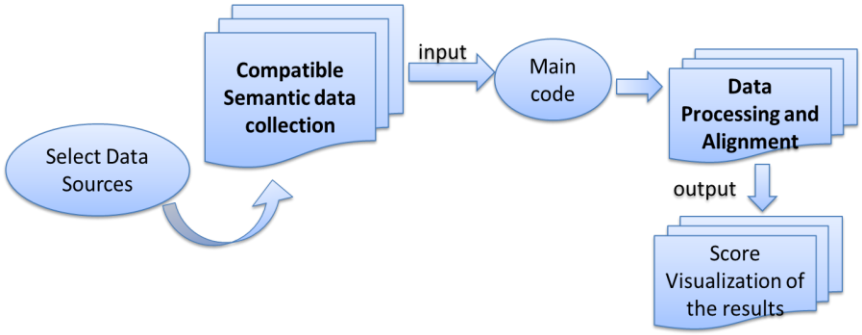


Figure 2 – Methodological Flowchart.

The first step was to identify and select data sources that were representative of the same mapping scale, contained semantic definitions of data classes, and fell within the scope of land use and cover. The chosen scale was 1:25,000, the smallest scale of national systematic mapping, allowing for generalization to the larger scales stipulated in the legislation (DSG, 2017).

To ensure the use of widely adopted standards, Brazil's national mapping agencies (DSG and IBGE) were the main sources of the national definitions for this study. These sources include the topographic mapping conceptual model ET-EDGV 3.0, homologated by the National Commission on Cartography (CONCAR), as well as the thematic models presented in the Land Use and Land Cover Manual and the Brazilian Vegetation Manual, both from IBGE. To complement these with global data sources, the Food and Agriculture Organisation's (FAO) Forest Resources Assessment and Dynamic World were also included (Brown et al., 2022; DSG, 2017; FRA, 2015; IBGE, 2012, 2013).

Some considerations should be made regarding the classes selected from the IBGE manuals to generate an input file for the main code. In the land use and land cover manual, the classes referring to anthropised areas, as this is not the purpose of the maps, were not considered in the input file, since this specific class, in topographic mapping, has a major complexity in terms of modelling and conceptualization of data, so a specific approach is suggested for these classes. The data classes related to water, and its uses were not considered for the same reason.

Classes related to vegetation transition systems were not considered, as a specific approach with the "Vegetação de contato" class on the topographic map is recommended. All remaining data class definitions were organized in a structured table, accessible via a link[1]. This table contains definitions for all classes and subclasses in the respective models. This table serves as our 'word box' and provides the basis for subsequent queries. This structured dataset ensures consistency and provides a centralized resource for comparing and aligning geospatial data classes across different sources.

2.3 Data Processing and Alignment

We utilised Python, a versatile and widely used programming language, to efficiently process and align data definitions. Python's well-established ecosystem facilitated seamless integration with OpenAI libraries, supporting a collaborative and reproducible research environment. The implementation was carried out using a Google Collaboratory Notebook, chosen for its accessibility and ability to host all the necessary natural language processing (NLP) tools employed in this study. The full code, divided into two primary sections (semantic search and similarity comparison), is publicly available at link[2], This repository includes comprehensive author notes and a streamlined version containing only the source code for simplified replication.

2.3.1 **Semantic Search Phase**

The first component of the methodology focuses on semantic search, which is crucial for identifying definitions with the highest semantic similarity to a given search term. The process is detailed in the pseudocode below:

```
Algorithm Semantic_Search Pseudocode

  INPUT: OpenAI_API_key, "words_PT1_frmt.csv"
  OUTPUT: Ranked list of definitions based on semantic similarity

  IMPORT necessary NLP libraries
  PROVIDE OpenAI_API_key to authenticate access

  LOAD "word_box_01.csv" containing definitions of land cover and land use classes
  COMPUTE embeddings for all definitions in "word_box_01.csv"

  PROMPT user to INPUT a search term (e.g., "forest")
  COMPUTE embedding for the search term

  FOR each definition in "word_box_01.csv"
    CALCULATE semantic similarity between search term embedding and definition embedding
  END FOR

  SORT definitions by descending order of similarity scores
  DISPLAY the ranked list of definitions
END Algorithm
```

[1] https://anonymous.4open.science/r/word_box-4177.

[2] https://anonymous.4open.science/r/MainCode-C56E.

This phase begins with importing the necessary NLP libraries and authenticating the application with an OpenAI API key. The data file, "word_box_01.csv," containing various land cover and land use class definitions, is then loaded and transformed into embedding vectors. A search term, such as "forest," is prompted from the user, and its embedding is computed. The similarity between the search term's embedding and each class definition's embedding is calculated, sorted, and displayed in descending order of similarity. This systematic approach identifies and ranks the most relevant definitions, facilitating semantic alignment across the dataset.

2.3.2 Similarity Between Class Definitions

The second phase of the methodology focuses on comparing the semantic similarity between all class definitions. This process is detailed in the following pseudocode:

```
Algorithm Similarity_Between_Classes Pseudo Code

  INPUT: "word_box_01.csv"
  OUTPUT: Similarity matrix and visual representations (heatmap and graphs)
  IMPORT necessary NLP and visualization libraries
  DEFINE transformation model for generating word embeddings

  LOAD data class definitions from "words.csv"
  COMPUTE embeddings for all class definitions

  INITIALIZE a zero matrix "sim" with dimensions (N x N), where N is the number of definitions

  FOR each pair of definitions (i, j) in the dataset
    CALCULATE cosine similarity between embedding of definition_i and definition_j
    STORE similarity value in "sim" matrix at position (i, j)
  END FOR

  DISPLAY "sim" matrix as a heatmap for visual inspection
  GENERATE relational graphs based on similarity values
END Algorithm
```

This phase generates a similarity matrix that maps the relationships between class definitions. Each entry in the matrix represents the cosine similarity score between pairs of class definitions, providing insights into the semantic alignment within the dataset. The output matrix is then visualized as a heatmap to enhance the interpretability of the semantic relationships. These visual tools facilitate a deeper analysis of how data classes relate semantically across different sources. These methodological steps outline a clear and reproducible framework for aligning geospatial data classes using advanced NLP and AI tools, allowing semantic interoperability in geospatial applications.

3. Results

The results of this study demonstrate the successful semantic alignment of definitions between Brazilian topographic mapping and other sources and the extraction of similarity metrics between mapped concepts. This section is divided into subsections to present the results and their implications.

3.1 Alignment of Definitions and Similarity Metrics

The output of the alignment process is a comprehensive table that can be accessed and downloaded via a provided link[3]. In a grey background, this table lists the class names from Brazilian topographic maps and their attributes in the first column, followed by their corresponding ET-EDGV definitions in the second column. The third to sixth columns include analogous definitions from other sources, chosen based on the highest similarity scores computed between the topographic map and definitions from each source.

This tabular layout visually represents semantic alignment between the topographic mapping definitions and those from other sources. Additionally, the last column includes a harmonized definition generated by ChatGPT-4.0. This approach highlights the potential of artificial intelligence to refine and enhance the clarity and completeness of geographic concept definitions. Table 1 exemplifies how definitions from different sources have been aligned, emphasizing the potential for automation and machine-readable processing in AI-based applications.

Table 1- Alignment Table of Data Classes with the highest scores between data definitions.

	ET-EDGV 3.0- DEFINITION	FRA – FAO DEFINITION	IBGE-Manual Uso e Cobertura- DEFINITION	IBGE-Manual Vegetação DEFINITION	Dynamic World - DEFINITION	CHAT GPT HARMONIZATION
Grassland / Campo	Campo é uma forma particular de ocorrência (normalmente circunstancial) de uma vegetação …/ Grassland is a particular form of occurrence (usually circumstantial) of vegetation …	Toda a terra que não seja classificada como floresta ou outra terra arborizada. /All land that is not classified as forest or other wooded land.	Entendem-se como áreas campestres as diferentes categorias de vegetação fisionomicamente bem diversa da florestal…/ Grassland areas are understood as the different categories of vegetation that are physiognomically very different from forest vegetation…	A Estepe Gramíneo-Lenhosa é o tipo mais representativo dos campos do sul do Brasil…/ The Grassy-Woody Steppe is the most representative type of grassland in southern Brazil…	Áreas abertas cobertas por gramíneas homogêneas com pouca ou nenhuma vegetação alta.../ Open areas covered by homogeneous grasses with little or no tall vegetation…	Campo é uma área de terra que não é classificada como floresta ou outra terra arborizada.../ Grassland is an area of land that is not classified as forest or other wooded land,
Clean Grassland / Campo Limpo	Vegetação predominantemente herbácea, com raros arbustos e ausência de árvores. /Predominantly herbaceous vegetation, with rare shrubs and no trees.	Toda a terra que não seja classificada como floresta ou outra terra arborizada./ All land that is not classified as forest or other wooded land.	Entende-se como áreas campestres as diferentes categorias de vegetação fisionomicamente bem diversa da florestal…/Grassland areas are understood as the different categories of vegetation that are physiognomically very different from forest vegetation…	A Estepe Gramíneo-Lenhosa é o tipo mais representativo dos campos do sul do Brasil… /The Grassy-Woody Steppe is the most representative type of grassland in southern Brazil…	Áreas abertas cobertas por gramíneas homogêneas com pouca ou nenhuma vegetação alta... /Open areas covered by homogeneous grasses with little or no tall vegetation…	O termo "Campo Gramíneo-lenhoso" refere-se a áreas de terra que não são classificadas como florestas ou terras arborizadas…/ The term "Grassland-Woodland" refers to areas of land that are not classified as forest or wooded land.

[3] https://anonymous.4open.science/r/Analogous-definitions_output-7362.

Dirty Grassland /Campo Sujo	Vegetação com fisionomia herbácea e arbustiva, com arbustos e subarbustos espaçados entre si…/ Vegetation with herbaceous and shrubby features, with shrubs and subshrubs spaced apart…	Terrenos definidos como “Outros terrenos Florestados ”, com mais de 0,5 hectares; com árvores com mais de 5 metros de altura e.../ Land defined as “Other Forested Land” over 0.5 hectares; with trees...	Entende-se como áreas campestres as diferentes categorias de vegetação fisionomicamente bem diversa da florestal, ou seja, aquelas que se caracterizam por …/ Countryside areas are understood as the different categories of vegetation that are physiognomically very different from forest vegetation, that is, those that …	As maiores extensões de Estepe Parque foram observadas na parte leste do Planalto das Araucárias, na porção central do Planalto …/ The largest extensions of Steppe Park were observed in the eastern part of the Araucárias Plateau, in the central portion of the Rio …	Áreas abertas cobertas por gramíneas homogêneas com pouca ou nenhuma vegetação alta. Outras áreas homogêneas de vegetação semelhante a gramíneas.../ Open areas covered by homogeneous grasses with little or no tall vegetation. Other homogeneous areas of grass-like vegetation…	O termo "Campo Parque" refere-se a terrenos que não são predominantement e florestados, agrícolas ou urbanos, com características específicas de vegetação.../ The term "Campo Parque" refers to land that is not predominantly forested, agricultural or urban, with specific vegetation characteristics.

3.2 Quantitative Analysis of Semantic Alignment

The degree of similarity was assessed by aligning semantic definitions from different sources with the Brazilian topographic mapping classes, represented as S = value. The highest similarity scores aligned with analogous classes, enabling the creation of alignment diagrams that map data model relationships. The data relation cardinalities, such as 1...* (one correspondence in one model to many in another) and *...1 (many correspondences in one model to 1 in another), illustrate the level of detail of the aligned data. For example, in the diagram comparing the ET-EDGV and the Brazilian Vegetation Manual, we observe that the forest classes in the topographic map are represented by a single class. In contrast, this concept is represented in vegetation mapping by six main classes, as illustrated in the diagram, and 26 more subclasses from these. All the definitions of the subclasses were inserted in the input file of the main code for all the models' classes. In this case of the forest cardinality 1...*, the highest score value was considered to represent it in the diagram. Figure 3 illustrates the semantic alignment between Brazilian topographic maps and national and international data sources. Note that the data class terms have been kept in their original languages.

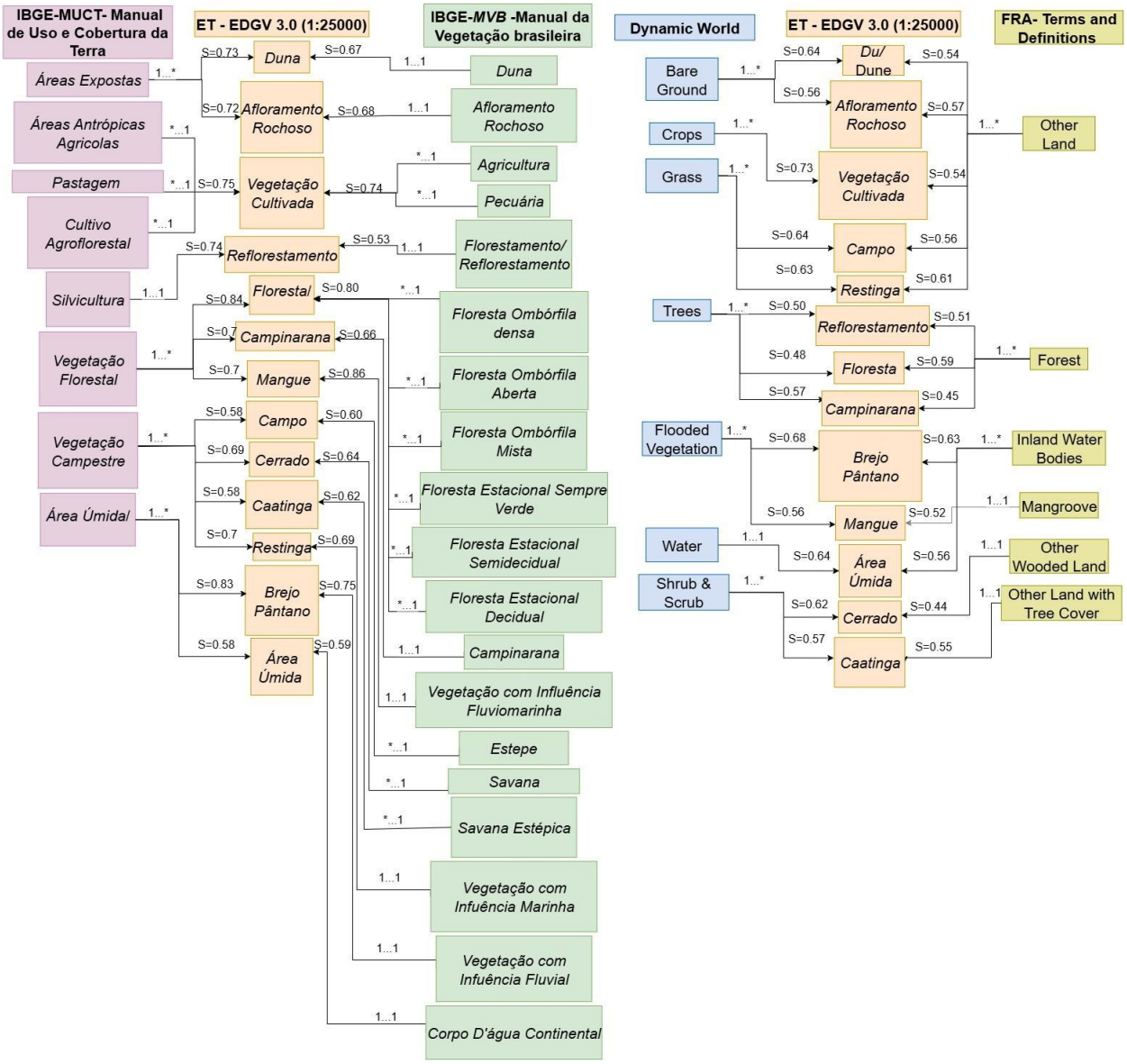


Figure 3 – Alignment Diagram between the Brazilian Topographic Map and other sources.

3.3 Correlation Analysis

A correlation matrix was constructed to analyze the semantic relationships between class definitions from different sources. This matrix represents the semantic similarity scores for all pairs of definitions, with values ranging from 0 (no similarity) to 1 (identical definitions). The matrix's main diagonal contains the value 1, as each definition is compared to itself. The matrix reveals how definitions from different sources closely align with the Brazilian topographic mapping standard (ET-EDGV 3.0). Table 2 shows a subset of these values, providing insights into the semantic alignment across datasets. Notably, higher similarity scores were observed for more detailed definitions from IBGE sources, while more generic.

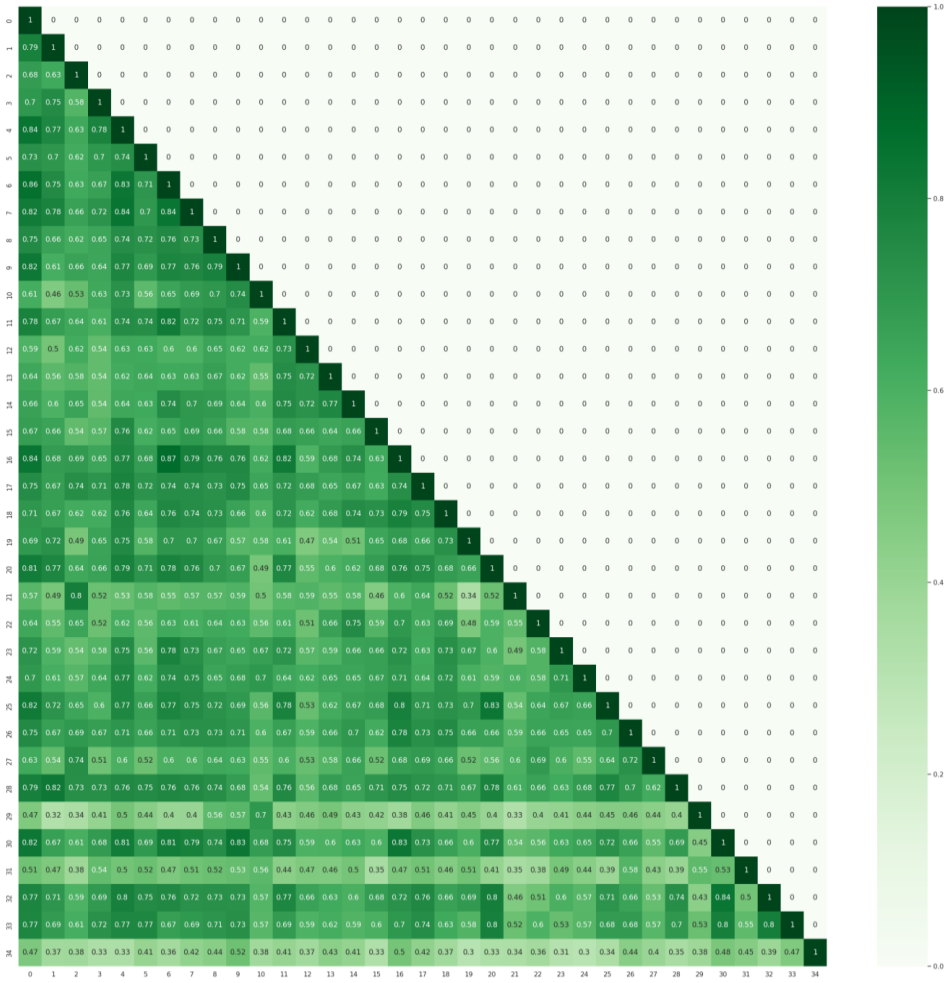
Table 2 - Aligned data class correlation values (top scored) and nomenclature.

ET-EDGV 3.0 Class	**FRA - FAO Class/SCORE**	**IBGE MUCT Class/SCORE**	**IBGE MVB Class/SCORE**	**Dynamic World Class/SCORE**
Grassland/ *campo*	Other Land/ *Outras Áreas*/**0.56**	Grassland Vegetation/ *Vegetação Campestre*/**0.58**	Steppe / *Estepe*/**0.6**	Grass/*Grama*/**0.64**
Cultivated Vegetation/ *Vegetação Cultivada*	Other Land/ *Outras Áreas*/**0.54**	Agricultural Area/*Área Agrícola*/**0.75**	Agriculture/ *Agricultura*/**0.74**	Crops/Colheita/ **0.73**
Mangrove/ *Mangue*	Mangrove/ *Mangue*/**0.52**	Forest Vegetation/ *Vegetação Florestal*/**0.7**	Fluvio Marine Influence Vegetation / *Vegetação com influência Fluviomarinha*/ **0.86**	Flooded Vegetation/ *Vegetação* Submersa/ **0.56**
Forest/ *Floresta*	Forest/ *Floresta/* **0.59**	Forest Vegetation/ *Vegetação Florestal*/ **0.84**	Dense rainforest / *Floresta Ombrófila Densa*/ 0.8	Trees/ Árvores/**0.48**
Wetland/ *Área Úmida*	In Land Water Bodies/*corpos de água terrestre/* **0.56**	Wetland / *Área Úmida* **0.58**	Continental Water Body/ *Corpos D'água Continentais*/**0.59**	Water/Água/**0.64**
Steppe Savannah /*Caatinga*	Other Land with Tree Cover/ *Outras terras com cobertura arbórea*/ **0.55**	Grassland Vegetation/*Vegetação Campestre*/**0.58**	Steppe Savannah/ *Savana Estépica/* **0.62**	Shrub & Scrub/ *Arbusto e Matagal* /**0.57**
Savannah /*Cerrado*	Other Wooded Land/ *Outras terras arborizada/* **0.44**	Grassland Vegetation/ *Vegetação Campestre* /**0.69**	Savannah/ *Savan/* **0.64**	Shrub & Scrub/ *Arbusto e Matagal* / **0.62**
Campinarana	Forest / *Floresta*/**0.45**	Forest Vegetation/ *Vegetação Florestal* /**0.7**	*Campinarana*/**0.66**	Trees/ Árvores/ **0.57**
Reforestation/ *Reflorestamento*	Forest / *Floresta*/**0.51**	Forestry/ *Reflorestamento*/**0.74**	Reforestation/ *Reflorestamento* /**0.53**	Trees/ Árvores/ **0.5**
Restinga	Other Land/ *Outra Terr/* **0.61**	Wetland / *Área Úmida* **0.7**	Marine Influence Vegetation / *Vegetação com influência Marinha/* **0.69**	Grass/ *Grama/* **0.63**
Marsh or Swamp/ *Brejo ou Pântano*	In Land Water Bodies/ *corpos de água terrestre*/**0.63**	Wetland / *Área Úmida* 0.83	Rriver Influence Vegetation/ *Vegetação com influência Fluvial*/**0.75**	Flooded Vegetation / *Vegetação* Submersa/**0.68**
Dune/*Duna*	Other Land/ *Outra Terra*/**0.54**	Exposed Areas/*Áreas Expostas*/**0.73**	Dune / *Duna/* 0.67	Bare Ground/ *Terra nua*/**0.66**
Rocky Outcrop/ *Afloramento Rochoso*	Other Land/ *Outra Terra*/**0.57**	Exposed Areas/ *Áreas Expostas*/**0.72**	Rocky Outcrop/ *Afloramento Rochoso*/ 0.68	Bare Ground/ *Terra nua*/**0.56**

3.4 Visualizing Results with Heat Maps

To better visualize these trends, a heat map was created to represent the correlations of each data element with every other element. This visualization helps to identify similarity between definitions from different sources. The average similarity between two classes can be calculated by grouping definitions from the same data class. For example, the similarity between all definitions of "Forest" across different sources was determined, allowing for further discussion on the usefulness of this coefficient. Figure 4 illustrates the heat map applied to these class definitions. It can be observed that greater generality definitions present low correlation values between the other definitions, resulting in a lighter tone line, as in lines 34 and 31. In contrast, high correlations occur through dark points or darker patches.

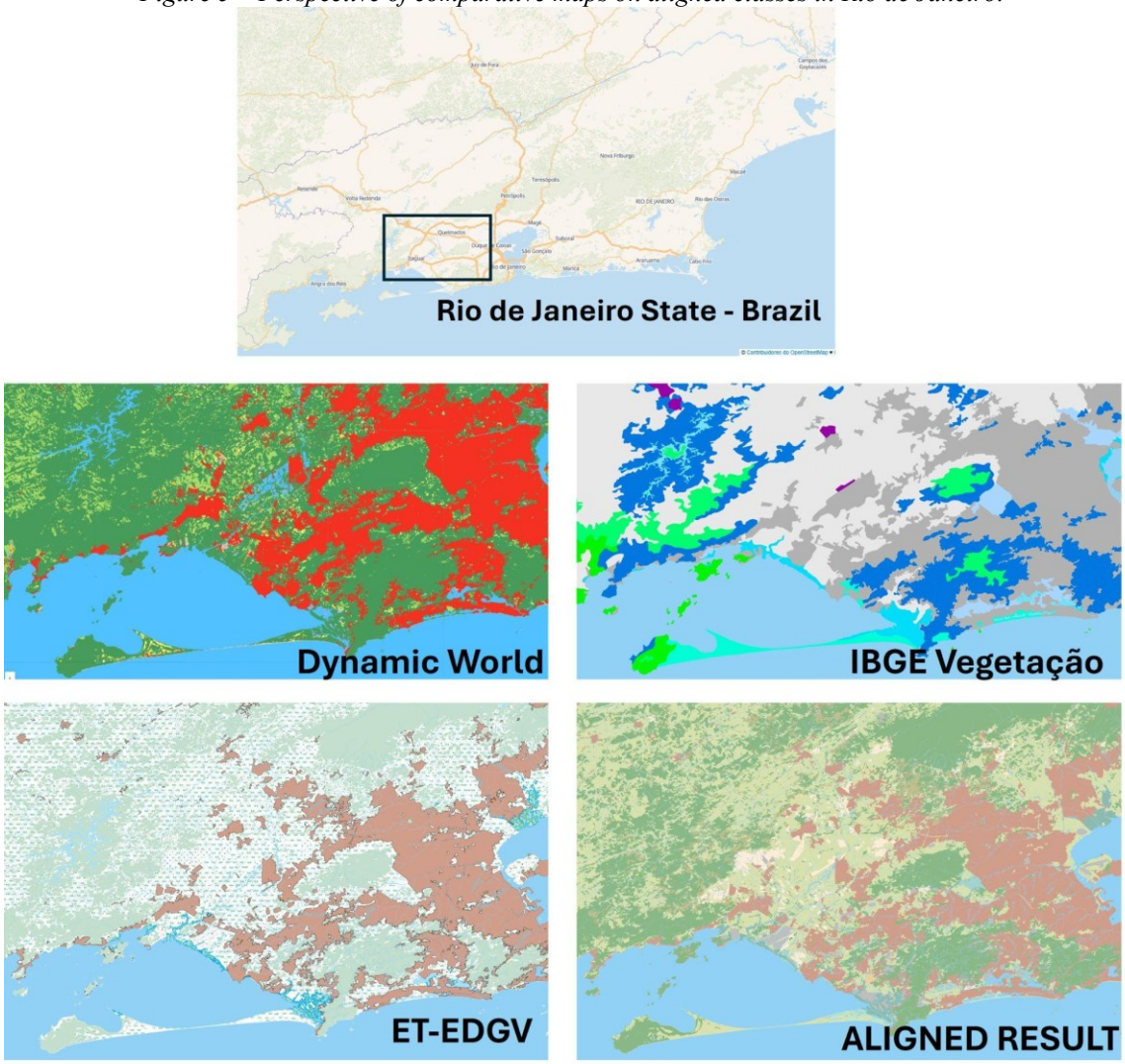
Figure 4 – Forest Heat Map with the Scores class by class.



Maps were generated to provide a comparative perspective on aligned classes within the same geographical area. The maps generated employed the aligned classes for each model, with the objective of achieving a uniform presentation. The selected region was part of the state of Rio de Janeiro, and all images were represented at an approximate scale of 1:250,000. The legends for the land use and land cover map and the topographic map are presented in a unified manner. The vegetation map was created in accordance with the regulations delineated in the pertinent technical map manual. The

legend was derived from the IBGE Environmental Information Bank, which can be accessed via the provided link[4]. The Dynamic World map, in combination with its associated legend, was retrieved from the project's official platform. The FAO vegetation map was excluded due to the lack of comparable data. Nevertheless, the comparative maps and their respective legends can be accessed via the following link[5].

Figure 5 – Perspective of comparative maps on aligned classes in Rio de Janeiro.



4. Discussion

This study has provided valuable insights into the alignment between different land use and vegetation classification systems, specifically the IBGE Land Use and Land Cover and Vegetation Handbook, the ET-EDGV Topographic Mapping, and international standards such as Dynamic World and the Global Forest Resources Assessments (FRA). These

[4] https://bdiaweb.ibge.gov.br/#/consulta/vegetacao

[5] https://anonymous.4open.science/r/Comparative_Maps-3247/README.md

insights help elucidate both the potential for integration and the challenges faced with diverse mapping standards. Artificial intelligence and NLP techniques, such as those implemented in ChatGPT-4.0, also generated harmonized definitions. These may represent a promising path for improving the clarity and consistency of individual classification systems, though they do not offer a definitive solution.

Integrating different land cover and land use (LULC) classification systems reveals the complexity of aligning datasets with varying levels of detail and thematic focus. The ET-EDGV, as a topographic mapping standard, provides a higher level of detail due to its reliance on aerial photography and field verification. This contrasts with global standards, which tend to be more generic and can encompass multiple ET-EDGV categories. Notably, the IBGE Brazilian Vegetation Handbook demonstrates a high degree of detail, often closely aligning with ET-EDGV classes, sometimes in a near 1:1 correspondence. Integrating detailed topographic mapping data into broader LULC datasets tends to be more semantically accurate than the reverse, although some information loss is inevitable. For example, the FRA's focus on forests illustrates semantic divergence arising from differing objectives and user needs.

A central contribution of this work is the use of artificial intelligence (AI) to quantify and support semantic connections between classification definitions. Applying similarity values (S-values) allows for the measurement of class alignment, often mirroring human-level interpretation. These values revealed strong alignments in cases such as "Cultivated Vegetation" and "Crops," highlighting the potential for seamless data integration. However, discrepancies emerged, such as those between "Grassland" in the IBGE LULC, "Reforestation" in the IBGE Vegetation Handbook, and ET-EDGV, indicating the need for further investigation. Unique regional classes, like "Campinarana" from the Amazon, exhibited low S-values compared to international classifications, reflecting the challenge of aligning region-specific types with broader, global categories. Similarly, discrepancies such as the low S-values for "Mangrove" between ET-EDGV and FRA suggest classification criteria or scope differences.

Another important finding relates to the concept of cardinality in semantic alignment. One-to-one (1:1) class relationships typically resulted in higher S-values and stronger semantic alignment than one-to-many (1...*) relationships. This underscores the greater ease of achieving semantic accuracy when classifications map directly rather than requiring aggregation or disaggregation. These challenges are especially evident when translating detailed national categories, like ET-EDGV's "Wetlands," into more generalized international classes, often requiring simplifications that reduce specificity and may lead to misalignments.

Despite these challenges, AI-assisted similarity scoring is valuable in identifying optimal alignments and promoting data interoperability. Future work could explore refining international standards to accommodate better region-specific ecosystems, such as "Campinarana," improving the representation of unique biomes. Investigating low S-values in greater depth can also help to fine-tune AI models for improved semantic matching. Efforts to enhance AI methodologies should focus on better handling one-to-many relationships and the nuanced characteristics of thematic data. Furthermore, adopting open-source large language models (LLMs), such as Llama, could reduce dependence on proprietary technologies like those from OpenAI.

5. Conclusion

Describing the landscape has long been an essential mapping function, especially for topographic representations. This study confirms that topographic mapping, characterized by high detail and accuracy, can effectively be aligned with broader LULC and thematic classifications using AI-driven methodologies. By employing NLP techniques and using semantic similarity as a key measure, this research addressed the challenge of aligning disparate geospatial data sources, contributing to improved interoperability and more robust integration practices. The methodology outlined demonstrated that AI can bridge semantic gaps, creating connections between data sources that mirror human-level understanding and aligning with the principles of geosemantics as described by Kuhn (2005). When used thoughtfully, the study showed that semantic similarity values could guide data harmonization, reducing manual effort and minimizing human bias while ensuring consistency and relevance across varied mapping frameworks. The main result is characterized by the alignment between data classes through semantic similarity between formal definitions. While the focus was on Brazilian geospatial

data—highlighting unique ecosystems such as “Campinarana”—the methodology has global applicability. It provides a scalable framework for integrating diverse geospatial datasets. For further investigations, it is recommended to adopt local factors, such as climatic characteristics, to minimize the specificities of each formation, especially among global models. Future work should incorporate evolving AI models and expand the method to include additional ecosystems and data types. Continued advancements in NLP and AI are expected to enhance the semantic precision of data integration, fostering a deeper understanding of landscape representations and their semantic properties.

References

Ali, A., Alfayez, F., & Alquhayz, H. (2018). Semantic Similarity Measures Between Words: A Brief Survey.

Andeson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). Professional Paper (Professional Paper). https://books.google.com.br/books?hl=en&lr=&id=dE-ToP4UpSIC&oi=fnd&pg=PA7&dq=+A+Land+Use+and+Land+Cover+Classification+System+for+&ots=sZki0-_l5E&sig=A72-24e0caZeg-TGQYRtmQjRMcc

Bennett, B. (2001). What is a Forest? On the Vagueness of Certain Geographic Concepts. https://doi.org/10.1023/A:1017965025666

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. 5. https://doi.org/10.48550/arXiv.1607.04606 , Focus to learn more.

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., & Tait, A. M. (2022). Dynamic World, Near real-time global 10 m land use land cover mapping. Scientific Data, 9(1), 251. https://doi.org/10.1038/s41597-022-01307-4

Chandrasekaran, D., & Mago, V. (2022). Evolution of Semantic Similarity—A Survey. ACM Computing Surveys, 54(2), 1–37. https://doi.org/10.1145/3440755

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). https://doi.org/10.18653/v1/N19-1423

Dharma, E. M., Gaol, F. L., Warnars, H. L. H. S., & Soewito, B. (2022). The Accuracy Comparison Among Word2vec, Glove, And Fasttext Towards Convolution Neural Network (Cnn) Text Classification.pdf. 100(2). https://doi.org/10.29207/resti.v6i3.3711

Doyle, J. F. (1973). Can Satellite Photography Contribute To Topographic Mapping? p. 315–325.

BRASIL. (2017). Especificações Técnicas Para Estruturação De Dados Geoespaciais Vetoriais (Et-Edgv 3.0). Ministério Do Planejamento, Desenvolvimento e Gestão Comissão Nacional de Cartografia; PDF. https://www.bdgex.eb.mil.br/portal/index.php?option=com_content&view=article&id=81&Itemid=353&lang=pt

Elavarasi, S. A., Akilandeswari, D. J., & Menaga, K. (2014). A Survey on Semantic Similarity Measure.

FRA. (2020). FRA 2020 Terms and Definitions. Viale delle Terme di Caracalla Rome 00153, Italy. https://www.fao.org/3/I8661EN/i8661en.pdf

Fremlin, G., & Robinson, A. H. (1998). What Is It That Is Represented on a Topographical Map? Cartographica: The International Journal for Geographic Information and Geovisualization, 35(1–2), 13–19. https://doi.org/10.3138/CP64-0LM7-0P51-PT77

Gersmehl, P. J. (1981). Maps In Landscape Interpretation. Cartographica: The International Journal for Geographic Information and Geovisualization, 18(2), 79–115. https://doi.org/10.3138/Q508-5316-U142-R6G3

Gorman, J., & Curran, J. R. (2006). Scaling distributional similarity to large corpora. Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL - ACL '06, 361–368. https://doi.org/10.3115/1220175.1220221

IBGE (Ed.). (2012). Manual técnico da vegetação brasileira (2a edição revista e ampliada). Instituto Brasileiro de Geografia e Estatística-IBGE. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=263011

IBGE (Ed.). (2013). Manual técnico de uso da terra (3a edição). Instituto Brasileiro de Geografia e Estatística-IBGE. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=232440

Kuhn, W. (2005). Geospatial Semantics: Why, of What, and How? In S. Spaccapietra & E. Zimányi (Eds.), Journal on Data Semantics III (Vol. 3534, pp. 1–24). Springer Berlin Heidelberg. https://doi.org/10.1007/11496168_1

Langran, G. (1985). Map Design for Computer Processing: Literature Review and DMA Product Critique.

Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings.

Machado-García, N., González-Ruiz, L., & Balmaseda-Espinosa, C. (2014). Recuperación de objetos geoespaciales utilizando medidas de similitud semántica. 8(2).

Mallenby, D. (2008). Handling Vagueness in Ontologies of Geographical Information. https://academiccommons.columbia.edu/doi/10.7916/D8D50KPG

Manning, C., Raghavan, P., & Schuetze, H. (2009). Introduction to Information Retrieval.

Martinez-Gil, J. (2022). A comprehensive review of stacking methods for semantic similarity measurement. Machine Learning with Applications, 10, 100423. https://doi.org/10.1016/j.mlwa.2022.100423

Meng, L., Huang, R., & Gu, J. (2013). A Review of Semantic Similarity Measures in WordNet. International Journal of Hybrid Information Technology, 6(1).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/D14-1162

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1), 17–30. https://doi.org/10.1109/21.24528

Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications, 39(9), 7718–7728. https://doi.org/10.1016/j.eswa.2012.01.082

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/D15-1036

Silva, L. S. L., & Camboim, S. P. (2020). Brazilian Nsdi Ten Years Later: Current Overview, New Challenges And Propositions For National Topographic Mapping. Boletim de Ciências Geodésicas, 26(4), e2020018. https://doi.org/10.1590/s1982-21702020000400018

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019). A Comparison of Semantic Similarity Methods for Maximum Human Interpretability. 2019 Artificial Intelligence for Transforming Business and Society (AITB), 1–4. https://doi.org/10.1109/AITB48515.2019.8947433

Varzi, A. C. (2001). Vagueness in geography. Philosophy & Geography, 4(1), 49–65. https://doi.org/10.1080/10903770124125

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Wang, Z., Mi, H., & Ittycheriah, A. (2017). Sentence Similarity Learning by Lexical Decomposition and Composition (arXiv:1602.07019). arXiv. http://arxiv.org/abs/1602.07019

Wilson, B. J., & Schakel, A. M. J. (2015). Controlled Experiments for Word Embeddings (arXiv:1510.02675). arXiv. http://arxiv.org/abs/1510.02675

Yuhua Li, Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering, 15(4), 871–882. https://doi.org/10.1109/TKDE.2003.1209005

Zhu, G., & Iglesias, C. A. (2017). Computing Semantic Similarity of Concepts in Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering, 29(1), 72–85. https://doi.org/10.1109/TKDE.2016.2610428