# Ontology extraction and evaluation for the Blue Amazon

Vivian Magri A. Soares[1,2], Renata Wassermann[1,2]

[1]Universidade de São Paulo, Instituto de Matemática e Estatística, Rua do Matao 1010, Cidade Universitaria, São Paulo, SP, Brazil
[2]Center for Artificial Intelligence (C4AI), Av. Prof. Lúcio Martins Rodrigues, 370, Cidade Universitaria, São Paulo, SP, Brazil

**Abstract**

The Brazilian maritime territory, often referred to as Blue Amazon, has invaluable significance for its resources, biodiversity, commercial importance etc. Yet, information about it is disperse. This project is geared towards the organization of knowledge on the form of an ontology. Searching for efficient methods with satisfying results for this task, a recent approach involving Large Language Models (LLMs) in the role of experts for building a conceptual hierarchy has shown promising results. This work presents the proposal for the experimentation with the construction of an ontology about the Blue Amazon related concepts using LLMs, followed by human and application-based evaluations.

**Keywords**

ontology extraction, ontology evaluation, Large Language Models, Blue Amazon

## 1. Introduction

The Brazilian maritime territory, a vast area with approximately the same size as the Amazon rainforest, and often referred to as Blue Amazon, is a region of invaluable importance, for Brazil and for the world, because of its economical and commercial resources, its multiple different ecosystems, and even for its key role in climate regulation. Yet, information about it is dispersed [1]. As such, it has become the focus of this project the organization of knowledge of this domain on the form of an ontology.

The manual construction of ontologies is an strenuous endeavor which requires access to domain experts. Many (semi-)automatic ontology extraction methods have been proposed over the years. Nevertheless, the problem of obtaining a well structured, relevant ontology without a fair amount of human labor persists. In that scenario, a recent approach [2] has shown promising results. It involves using a Large Language Model (LLM) in the role of experts providing subconcepts of a seed concept iteratively. This work is still preliminary and yields but a simple ontology, with only "is-a" type of relation. We believe, however, that this methodology has great potential in the aid of ontology learning, with plenty room for expansion over future works.

To assess how successful the LLM-powered algorithm is on the task, we envisioned two kinds of evaluation for the outputs. The first aided by human domain experts; and, then, testing the efficiency of the validated ontologies in enhancing the accuracy of a conversational agent.

This work contains the discussion of the concepts grounding our project, as well as the context of its development, and presents the current results and the proposal for the next steps of this research.

## 2. Related Work

Various techniques from different fields have contributed for the improvement of ontology development and extraction, and in the first part of this section we will highlight some of them. In the second part, we discuss some measures for ontology evaluation.

*Proceedings of the 17th Seminar on Ontology Research in Brazil (ONTOBRAS 2024) and 8th Doctoral and Masters Consortium on Ontologies (WTDO 2024), Vitória, Brazil, October 07-10, 2024.*

vivian.magri@ime.usp.br (V. M. A. Soares); renata@ime.usp.br (R. Wassermann)
https://www.ime.usp.br/~renata (R. Wassermann)
0009-0009-9767-4127 (V. M. A. Soares); 0000-0001-8065-1433 (R. Wassermann)

## 2.1. Ontology Extraction

Asim et al. [3] state ontologies can be created by extracting information from unstructured text in a step-by-step process known as ontology learning layer cake. In this approach, after preprocessing text corpora using linguistic techniques, relevant terms and concepts of a domain are extracted utilizing techniques of natural language processing (NLP). These methods may also be applied to obtain taxonomic and non-taxonomic relations among these concepts. The authors still mention semantic lexicons, which can be used at both term/concept extraction and relationship extraction stage, and ILP, that may be used to form axioms in the later stages [3].

More recently, though, we observed a great rise on the use of deep neural network-based methods. Reshadat et al. [4] say deep neural networks are powerful approaches for ontology population since the feature engineering procedure is done automatically. Furthermore, these systems are not constrained to a predefined set of relations and can extract any type of relation from a massive and unstructured corpus automatically. The disadvantages, however, are that usually these approaches either require an annotated corpus with concepts and the relations between them [4], or they deflect from the more rigid framework that constitutes an ontology.

Analyzing the literature, we observe that the purely automatic information extraction systems using the aforementioned approaches usually have trouble keeping the information at the same time complete, coherently hierarchical and within the domain, and at the useful level of granularity and relevance for specific applications. This is probably the reason why rule-based methods have been common for the task of ontology population [4]. One great inconvenience, though, is that these methodologies basically require the schema of the ontology to be chosen in advance and then provided as an input, and designing a schema for an entire domain is a non-trivial task itself that requires a domain expert and involves many design decisions [2]. Another way of dealing with this issue lies on the proposal of various approaches to semi-automatic ontology construction. But, while such approaches look good on paper, there's little evidence they have been applied in practice [2].

In this context, Funk et al. [2] presented an approach to use OpenAI's GPT 3.5 API to aid on the construction of a concept hierarchy for a context provided by the user. GPT stands for Generative Pre-trained Transformer, a type of Large Language Model (LLM). Over the past decade, advancements in natural language processing and machine learning have led to the development of increasingly sophisticated Language Models [5]. The rapid development of (what became known as) LLMs in recent years has also been fueled by growth in computational resources, availability of large datasets and evolving software stacks [6]. Rozière et al. [7] claim LLMs have reached a proficiency in natural language that allows them to be commanded and prompted to perform a variety of tasks, including ones that require advanced natural language understanding.

The general strategy in [2] is to take as input from the user a seed concept (for example, Animals, Music, or even Things) that will determine the domain. The algorithm then provides a textual description for it, identifies its subconcepts, and places it on the hierarchy being constructed. And so the loop continues with the exploration of every subconcept discovered, until the model either can not find any more new subconcepts, or some stop condition defined by the hyperparameters, like the maximum exploration depth, is met. The outputs include an owl file, with all concepts and their definitions, and a graph that allows for the visualization of the produced hierarchy. A small example of the latter can be seen on figure 1. Although the authors recognize further investigation is necessary to draw general conclusions, they believe their experiments indicate that LLMs can be of considerable help for constructing concept hierarchies and that the research has great potential for expansion.

## 2.2. Evaluation of Ontologies

The literature shows that there is no consensus regarding the classification of ontology evaluation proposals [8]. The diversity of initiatives poses difficulties to creating a unified classification [9]. One possible grouping for the techniques are the four categories proposed by Asim et al. [3]: a) Golden standard-based evaluation, that evaluates resultant ontology with a predefined benchmark, which

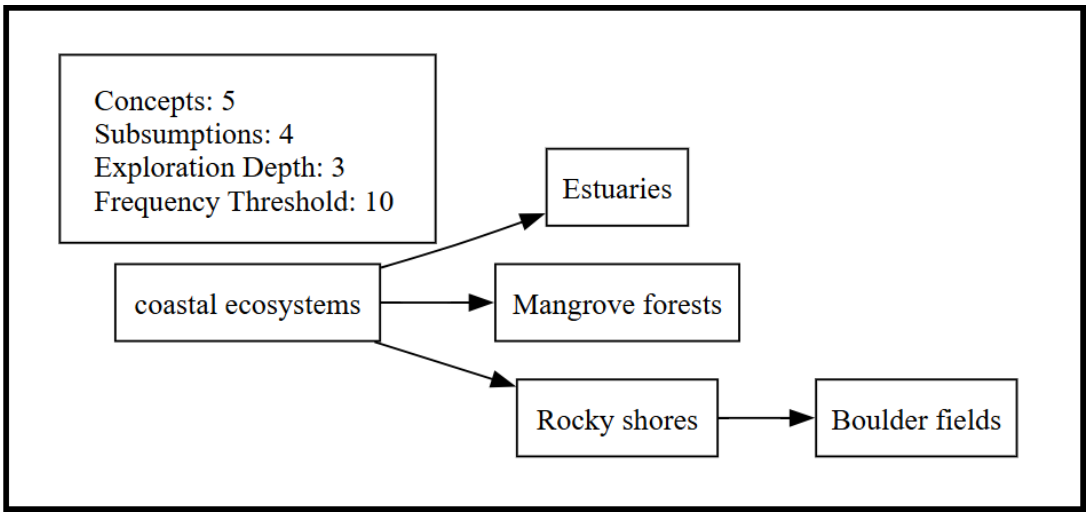


**Figure 1:** Example of output of the algorithm from [2] using "coastal ecosystems" as input

depicts an ideal ontology of a particular domain; b) Application-based evaluation, also referred as 'Task Based Evaluation', that evaluates a given ontology by exploiting it in a specific application to perform some task. The outcome of particular task determines the goodness of the specified ontology regardless of its structural properties; c) Data-driven evaluation (or so-called Corpus-based evaluation), that utilizes existing domain-specific knowledge sources (usually textual corpora) to assess the extent of coverage by one or more target ontologies in a particular domain; and d) Human evaluation, also called 'Criteria Based Evaluation', which is generally based on defining various decision criteria for the selection of the best ontology from a specified set of candidate ontologies. A numerical score is assigned as experts rate each relevant aspect of an ontology and a weighted sum is calculated.

As a technique that fits precisely the above definition of human evaluation, we can highlight ONTOMETRIC [10]. In order to solve what they call the election problem, when many candidate ontologies are available, they present a taxonomy of 160 characteristics that provides the outline to compare them. It is the skeleton used to build the multilevel tree of characteristics, that should be adapted by adding or removing characteristics according to their relevance for a given application of this method. The superior level of the taxonomy originally possesses five dimensions, i.e., the main aspects that the user should consider to examine an ontology. These are: content; language of implementation; methodology used for the development; software tools used for building it; and costs for the ontology in a certain project. Each dimension is defined through a set of factors, that is, the fundamental elements that should be analyzed to obtain the value of the dimensions. The factors, in turn, are defined through a group of characteristics that allow calculating the value of their suitability. These characteristics can be defined, recurrently, by means of even more specific subcharacteristics. Although the specialization of the characteristics and the assessment of the criteria of a particular ontology require considerable effort, feedback from project managers reveals that once the framework has been defined, and if it is applied to one particular type of ontology, ONTOMETRIC helps to justify decisions taken and to weigh up the choice of one ontology from other options [10].

Not every human evaluation method, however, is aimed at comparing ontologies. Almeida [9] proposes a multidisciplinary approach to ontology content evaluation using experts. Their research was conducted with the use of a simple search engine developed to allow the visualization of an ontology by the user, and a group of three questionnaires to be answered next. Those were prepared using concepts from distinct research fields that are related to content assessment. The questionnaire related to Information Quality was made based on criteria such as coverage, accuracy and content. The one based on Competency Questions means to assess the capacity of an ontology-based system to answer the questions it was designed to address. Finally, the questions based on educational objectives are intended to assess whether specific content was learned by a person during the exploration of the

ontology. The study concluded it is a good idea to aid the domain experts in verifying the quality of the conceptualization present in the ontology according to scientific criteria in order to validate the adequacy of the specification of a model [9].

Human evaluation have the possibility of covering all high levels of evaluation for ontologies distinguished on the [3] review, that is, Lexical, vocabulary, concept and data; Hierarchy and taxonomy; Other semantic relations; Context and application; Syntactic and Structure, architecture and design. Their major shortcoming is the requirement of high manual cost in terms of time and effort. Araújo and Lima [8] also point out the difficulties in establishing who the right users are and what the best criteria are for evaluating the ontology. Furthermore, the fact that ultimately the evaluation still depends on the expert's intuition makes it uncertain whether it is accurate. However, they state there is criticisms in the literature regarding each of the types of proposals for ontology evaluation, and that it is necessary to take into account what are the objectives of the evaluation and what exactly is intended to be evaluated in the ontology.

## 3. Context

Approximately as big as the Amazon rainforest, the Brazilian maritime territory is often referred to as the Blue Amazon [1]. In total, the it amounts to around 45,000,000 km² of sea. The Blue Amazon carries 95% of Brazil's international trade, 90% of its oil reserves and 77% of the country's gas reserves [1]. Another noteworthy aspect is its rich environment and biodiversity. The territory encompasses multiple different ecosystems [11]. The region is also a vital source of food supply and a key player in climate regulation [1]. Despite its importance, the Blue Amazon is still poorly documented.

Such was the context that originated the Knowledge Enhanced Machine Learning (KEML)[1], a research group integrated into the Center for Artificial Intelligence (C4AI)[2], which is a large research center headquartered at the Universidade de São Paulo, that congregates researchers and students from a wide variety of fields. KEML's objective has been to merge data-driven learning with knowledge-based reasoning [12]. [1] describes the group's first project, the BLue Amazon Brain (BLAB). BLAB was born with the goal of building an architecture aimed at disseminating information about the Brazilian coast domain and its importance [13]. It would function as a tool for education and environmental awareness.

Currently, the KEML team is mainly dedicated to producing conversational agents or related systems (such as QA Systems) based on LLMs, specially those enriched by Knowledge Representation (KR) mechanisms [13]. Among them is Blabinha 1.0 [14], a conversational agent specifically designed as an evaluation environment for LLMs and prompt engineering when placed in the role of conducting task-oriented and domain-oriented dialogues. Blabinha 1.0 is implemented using GPT-family models and prompt engineering of the chain-of-thought (step-by-step) type, aiming at promoting a child's engagement in a conversation about the Blue Amazon domain through a gamification strategy. During the conversation, the language model is subjected to a series of tasks, ranging from introducing itself to the child, to performing topic analysis and suggestion of subjects within the context [14].

## 4. Research Proposal

As shown in the previous sections, the Blue Amazon is a region of great importance for Brazil, about which structured information is scarce. Having such data organized and made available as an ontology could improve information distribution on the theme, as well as improve the maintenance of the data and it's integration on systems and in many AI applications. But the manual crafting of ontologies is a difficult engineering task that is both time consuming and costly [2]. Besides, it demands reasonable dominance on the subject. With that in mind, the previously discussed work [2] using an LLM to construct ontologies made of subconcept/is-a relations presented itself as a promising path for (semi)automatically extracting structured information about the Blue Amazon domain.

[1] https://sites.usp.br/keml/en/keml-en/
[2] https://c4ai.inova.usp.br/

The framework presented on their paper, however, is still recent, as is the use of LLMs in the construction of knowledge-based structures, and haven't yet been formally accessed. Thus, we consider relevant as a research goal to construct and conduct an evaluation for the results produced by their algorithm when applied in the domain of the Brazilian Coast. The evaluation will also be an opportunity to improve the outputs, as well as fine-tuning the workflow. It shall happen in two steps, using two different techniques, starting by employing the aid of experts on the domain to supervise the results, assessing the correctness and the design of the resultant ontologies. Finally, we intend to proceed with an application-based evaluation, using Blabinha 1.0 as a test environment.

We believe the contributions of this research will not only be the ontologies produced and evaluated, but also the findings to guide the use of the algorithm for the construction of concepts hierarchies and methods for their validation, as well as expanding the investigation on the LLMs potential for the construction of ontologies.

## 5. Preliminary Result

Currently, we have concluded the intended ontology generation part, and we are conducting the first step of the tests, working with the domain experts for the human validation of the outputs.

After installing the requirements and testing the code[3] described in [2] connected to the GPT 3.5 API, a few experiments with examples also used by the authors were performed, just to confirm the installation was functional. Then the concept "Blue Amazon" was first directly tested. The execution, however, was finished without obtaining any verified sub-concepts for it. As a workaround, "Brazilian Water Resources", a more general named concept, but still related to the original theme, was chosen for the experimentation with the main set of hyperparameters for the algorithm: Exploration depth (d) - up to this depth new concepts will be explored. The depth of a concept is its shortest path to the seed concept; and Frequency threshold (f) - lower values like 5 favor completeness, while higher values like 20 tend to benefit correctness. In accordance with the author's direct recommendation, two values suitable for test runs were used for the first, 2 and 3; and the range from 5 to 20 recommended for the second was covered using a step of 5. Thus, eight executions, all the possible combinations of the selected values, were tested using "Brazilian Water Resources" as the initial concept. From the observation of the results, we concluded that the combination of d = 3 and f = 10 produced the most interesting outputs. Therefore, these were the chosen values to be used in the subsequent experiments.

For the second round of tests, the chosen concept was "Coastal ecosystems". Since the initial tests showed the results where quite sensible to variations on the form of the input, a few different forms were experimented: first letter of the central concept capitalized or not (keeping the rest in lowercase in both cases, as the concepts outputted by the algorithm); referencing Brazil or not; concept in English or in Brazilian Portuguese (PT-BR). Also, the variations in Portuguese were used as input in the prompts modified to request the answers to be in PT-BR, as well. Finally, one concept was executed twice in the same conditions, to be considered as a parameter of the model's normal variation. In total, 13 executions, yielding ontologies containing between 1 and 36 concepts each, are being considered for analysis. The model struggled particularly to expand concepts in English referencing Brazil, such as "Brazilian Coastal ecosystems". Besides the edition to the prompts to force the PT-BR outputs, they were also slightly modified to reproduce the tests using GPT-4. These executions provided hierarchies containing between 2 and 99 concepts. This model had notably worse results when the version requesting answers in PT-BR (with instructions kept in English) was run.

For a deep evaluation of the outputs, it was considered essential to seek the aid of experts on the domain to oversee the results. We are employing five academics to respond forms aimed at evaluating each of the ontologies with at least 5 concepts. These experts have backgrounds related to the study of the Ocean, from fields such as Geosciences, Environmental Resource Management and Oceanography, most with experience in projects or studies related to Sustainability. Although five is not exactly a high number of people, the diversification of background should help diminish the bias of individual evaluation.

[3] https://git.informatik.uni-leipzig.de/hosemann/onto-llm

There is also a diverse range of age, gender, and educational levels — that go from undergraduate to post-doctor (some of which are also instructors).

The questionnaires were formulated mostly based on the work of [9] and [10]. From [10], the analysis of some factors, namely concepts, relations and taxonomy, from the dimension content, served as inspiration. Their work also motivated the idea of grouping the topics under evaluation hierarchically and the preparation for the calculation of a numeric score for each ontology. From [9], the main inspirations were the criteria related to information quality and the insights of how to use the questions to assess how well knowledge on the domain was being transmitted and to what degree an ontology was succeeding in the goal of modeling the real world concept. The higher level dimensions of our questionnaires are: accuracy, relevance, coverage, precision and information design. All of them, except for the last one, need to be analyzed at the concept level, considering each of their relationships. On this first round of evaluation it was considered important to go into this level of detailing to have a parameter over GPT's suggestions for each element of a hierarchy. The form also gives the opportunity for the respondents to express some impressions that will not be translated into a quantitative metric, but will be qualitatively analyzed to compose the results.

A first evaluation round was conducted, preparing forms for two smaller outputs, which were replied by the evaluators after a basic explanation, followed by a feedback meeting to both clarify their doubts, and to collect their insights to improve the evaluation. One important reminder that came from this experience is that it is incoherent to attempt an ontology evaluation without stating a clear purpose for it. Thus, the assessment is now being performed considering the goal of the ontologies as to construct a consistent concept hierarchy for the root concept of the given ontology considering what an elementary school student should be taught about such topic (in accordance with the next evaluation planned for the next steps, using Blabinha 1.0). The model form on its revised version, with the layout for the questions for each kind of concept (considering the type of relations it holds on a given ontology), can be found in: https://forms.gle/mxHiX61zUvu5BiqK6. Given the size of the outputs produced by GPT-4 (many exceeding 50 concepts), the model of the forms was simplified for the evaluation of these outputs. Great effort was put in to minimize the impact of this modification, as to guarantee it would only make it more practical, but the quality of the assessment for each dimension would be maintained. The latest model of the form is available on the link: https://forms.gle/oAbSjy6UnfKXgiFR6.

## 6. Next Steps and Future Works

The following activities are planned for this project, to be concluded, ideally, by the first trimester of 2025:

1. **Process the human evaluation results**
   After the conclusion of the specialists activities with the prepared forms, we must collect the answers, standardize the scales and calculate the metrics, as well as analyze the qualitative results. The metrics and the experts' impressions should provide means for us to compare the results between the different forms of inputs and the different models experimented, and hopefully derive enlightening conclusions.
2. **Perform the application-based evaluation**
   Following the assessing of the constructed ontologies, we intend to chose some of the outputs that achieve the best scores to merge into the Blabinha 1.0 prompts to test their impact on the model's performance. The aspect of topic analysis had previously been rated by human evaluation of the tool [4]. So, to measure the effect of inserting the concept hierarchies into the algorithm, a new round of evaluation would be conducted in a similar format to the previous one, focused on Blabinha's capacity to explore the domain of Coastal Ecosystems, in order to compare the results regarding the model's discernment about context.

[4]The results of this work are in process of publication.

We believe this research also has great potential for expansion beyond the scope of this project. For instance, through the construction of more ontologies, either using other concepts related to the Blue Amazon domain using the refinements to the pipeline our findings will provide, or applying the workflow for different contexts. As suggested by the article [2] itself, other relevant direction for future works are experimenting with the construction of ontologies that are more expressive, adding other kinds of relations and, possibly, even rules, as disjointness, for example; and testing the effect of fine-tuning for domain-specific ontology construction, training the model with curated information about the intended subjects.

## Acknowledgments

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with the support of the University of São Paulo, the São Paulo Research Foundation (FAPESP, grant #2019/07665-4) and by the IBM Corporation. Vivian Magri A. Soares was also supported by CAPES.

## References

[1] P. Pirozelli, A. B. R. Castro, A. L. C. de Oliveira, A. S. Oliveira, F. N. Cação, I. C. Silveira, J. G. M. Campos, L. C. Motheo, L. F. Figueiredo, L. F. A. O. Pellicer, M. A. José, M. M. José, P. de M. Ligabue, R. S. Grava, R. M. Tavares, V. B. Matos, Y. V. Sym, A. H. R. Costa, A. A. F. Brandão, D. D. Mauá, F. G. Cozman, S. M. Peres, The blue amazon brain (BLAB): A modular architecture of services about the brazilian maritime territory, CoRR abs/2209.07928 (2022). URL: https://doi.org/10.48550/arXiv.2209.07928. doi:10.48550/ARXIV.2209.07928. arXiv:2209.07928.

[2] M. Funk, S. Hosemann, J. C. Jung, C. Lutz, Towards ontology construction with language models, CoRR abs/2309.09898 (2023). URL: https://doi.org/10.48550/arXiv.2309.09898. doi:10.48550/ARXIV.2309.09898. arXiv:2309.09898.

[3] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, H. M. Abbasi, A survey of ontology learning techniques and applications, Database J. Biol. Databases Curation 2018 (2018) bay101. URL: https://doi.org/10.1093/database/bay101. doi:10.1093/DATABASE/BAY101.

[4] V. Reshadat, A. Akcay, K. Zervanou, Y. Zhang, E. de Jong, SCRE: special cargo relation extraction using representation learning, Neural Comput. Appl. 35 (2023) 18783–18801. URL: https://doi.org/10.1007/s00521-023-08704-9. doi:10.1007/S00521-023-08704-9.

[5] D. Nunes, R. Primi, R. Pires, R. de Alencar Lotufo, R. F. Nogueira, Evaluating GPT-3.5 and GPT-4 models on brazilian university admission exams, CoRR abs/2303.17003 (2023). URL: https://doi.org/10.48550/arXiv.2303.17003. doi:10.48550/ARXIV.2303.17003. arXiv:2303.17003.

[6] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zheng, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, B. Catanzaro, Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model, CoRR abs/2201.11990 (2022). URL: https://arxiv.org/abs/2201.11990. arXiv:2201.11990.

[7] B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bitton, M. Bhatt, C. C. Ferrer, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, G. Synnaeve, Code Llama: Open foundation models for code, 2023. arXiv:2308.12950.

[8] W. Araújo, G. Lima, O cenário da avaliação de ontologias: revisão de literatura, Tendencias da Pesquisa Brasileira em Ciência da Informação 9 (2016).

[9] M. B. Almeida, A proposal to evaluate ontology content, Appl. Ontology 4 (2009) 245–265. URL: https://doi.org/10.3233/AO-2009-0070. doi:10.3233/AO-2009-0070.

[10] A. L. Tello, A. Gómez-Pérez, ONTOMETRIC: A method to choose the appropriate ontology, J. Database Manag. 15 (2004) 1–18. URL: https://doi.org/10.4018/jdm.2004040101. doi:10.4018/JDM.2004040101.

[11] P. de M. Ligabue, A. A. F. Brandão, S. M. Peres, F. G. Cozman, P. Pirozelli, Blabkg: a knowledge graph for the blue amazon, in: P. Li, K. Yu, N. V. Chawla, R. Feldman, Q. Li, X. Wu (Eds.), IEEE International Conference on Knowledge Graph, ICKG 2022, Orlando, FL, USA, November 30 - Dec. 1, 2022, IEEE, 2022, pp. 164–171. URL: https://doi.org/10.1109/ICKG55886.2022.00028. doi:10.1109/ICKG55886.2022.00028.
[12] KEML, About keml, https://sites.usp.br/keml/en/keml-en/, 2023. Accessed: 2024-08-01.
[13] KEML, Conversational agents, https://sites.usp.br/keml/en/conversational-agents/, 2023. Accessed: 2024-08-01.
[14] KEML, Frameworks, https://sites.usp.br/keml/en/frameworks-2/, 2023. Accessed: 2024-08-01.