A machine learning domain ontology to populate knowledge base to support intelligent agents working in autonomous systems domains of the Internet infrastructure

Juliao Braga^{1,3}, Francisco Regateiro², Joaquim L. R. Dias², Itana Stiubiener³

¹Collaborating Researcher

²Instituto Superior Técnico, Universidade de Lisboa Lisbon, PT

> ³Universidade Federal do ABC Santo Andre, SP - Brazil

juliao.braga@ufabc.edu.br, francisco.regateiro@ist.utl.pt

joaquim.dias@tecnico.ulisboa.pt, itana.stiubiener@ufabc.edu.br

Abstract. This paper describes the creation of a domain ontology to represent knowledge to populate a knowledge base to be used by agents, in the environment of Internet Infrastructure routing domains. Protégé 5 was used, which produces results suitable for both software-developed agents and humans. The knowledge created with Protégé is explicit and Protégé has itself inference machines capable of producing implicit knowledge. The resources available in Protégé 5 are presented and the ontology is made available for public use. The content produced with Protégé 5 will be used to populate the knowledge base of the Structure for Knowledge Acquisition, Use, Learning and Collaboration (SKAU), an environment to support intelligent agents over Internet Autonomous Systems domains.

Keywords: internet infrastructure, autonomous systems, ontology, knowledge base, intelligent agents

Resumo. Este artigo descreve a criação de uma ontologia de domínio para representar conhecimento para popular uma base de conhecimento a ser usada por agentes, no ambiente dos domínios de roteamento da Infraestrutura da Internet. Foi utilizado o Protégé 5, que produz resultados adequados tanto para agentes desenvolvidos por software como para seres humanos. O conhecimento criado pelo Protégé é explícito e o próprio Protégé tem máquinas de inferência capazes de produzir conhecimento implícito. Os recursos disponíveis no Protégé 5 são exibidos e a ontologia é disponibilizada para uso público, em todas as suas versões. O conteúdo produzido pelo Protégé 5 será utilizado para povoar a base de conhecimento da Estrutura de Aquisição, Uso, Aprendizagem e Colaboração de Conhecimento (SKAU), um ambiente para apoiar agentes inteligentes nos domínios de Sistemas Autônomos da Internet. **Palavras-chave:** infraestrutura da internet, sistemas autonomos, base de conhecimento, agentes inteligentes, ontologia

1. Introduction

The goal of this paper is to describe the use of ontologies and issues associated with their production, to populate a knowledge base, usable by autonomous agents present in Internet Infrastructure routing domains (BRAGA et al., 2020). Ontology was defined in 1993 by Gruber and in 1997 this definition was adjusted more appropriately by Borst (GUARINO; OBERLE; STAAB, 2009) apud (GRUBER, 1993; BORST, 1997). In 1998, Studer and others adjusted the two definitions in the following definition proposal: "An ontology is a formal, explicit specification of a shared conceptualization" (STUDER; BENJAMINS; FENSEL, 1998) (GUARINO; OBERLE; STAAB, 2009). Several authors have refined the definitions over time to indicate, more clearly, that an ontology is a formally-defined vocabulary for a particular domain of interest used to capture knowledge about that (restricted) domain of interest. So, an ontology describes the concepts in the domain and also the relationships that hold between those concepts (HORRIDGE, 2011). The inclusion of a domain in the definition made it easier to understand the meaning of ontology. In this way and finally, it was defined ontology as describing concepts and relations in a specific domain are necessary for Knowledge Representation (KR) and Knowledge Exchange (KE) (SLIMANI, 2015).

The present work considers, as use case, a specific ontology for a repository of explicit knowledge on a well-defined machine learning domain represented by specifications that allow to extend this repository with implicit knowledge obtained through the action of reasoning engines (ANTONIOU; HARMELEN, 2004). This knowledge base is called Machine Learning Ontology (MLOnto) and can be used by humans and software-based agents (agents). The agents, in the context of this work, are part of the model named Autonomous Architecture Over Restricted Domains (A2RD) that depend on a knowledge base, one of the components that are hosted in a support environment named Structure for Knowledge Acquisition, Use, Learning and Collaboration (SKAU) (BRAGA et al., 2019). The SKAU ecosystem is visible in Figure 1.

MLOnto is built using Protégé 5 (Protégé), an ontology editor, which uses the Web Ontology Language (OWL) which, in turn, adopts Manchester syntax¹ (MUSEN, 2015) (HITZLER et al., 2009). Among the various languages available to represent knowledge, OWL brings together a set of favorable facilities because it has a high degree of expressiveness (Figure 2) through a reasonably compact syntax, precise semantics under its formalism and easy to be analyzed by reasoning mechanisms (KALIBATIENE; VASILECAS, 2011) (GOLBREICH et al., 2003). Protégé was created at Stanford University and is very popular in the field of Semantic Web. Protégé is free, developed in Java and its source code is released under a free license (the Mozilla Public License). Protégé can read and save ontologies in most ontologies formats: RDF, RDFS, OWL, etc. It is recognized for its ability to work on large Ontologies (MUSEN, 2015) (SLI-MANI, 2015) (TUDORACHE et al., 2013).

This work is a substantial update to the preliminary studies described in Braga, Dias e Regateiro (2020), and is divided into six sections, including this Introduction. Section 2 describes the related work. Section 3 describes the MLOnto architecture by identifying each of its seven highest-level classes. Section 4 describes its current implementation structure at Protégé. Section 5 establishes the conclusions and future work and the last

¹<https://www.w3.org/TR/owl2-manchester-syntax/>



Figure 1. The SKAU environment, support for agents of the A2RD model, where the KB displays the position of MLOnto and other ontologies. The other items that appear in the image of this figure are components that complete the SKAU environment, although important, are outside the scope of this paper. Adapted from (BRAGA, 2019b).



Figure 2. Languages Expressiveness. Source: (PEASE, 2011)

section 6 informs about the availability of the latest version of the ontology database.

2. Related Work

The use of Machine Learning to build Ontologies or a subset of Ontologies has been common in the literature (POON; DOMINGOS, 2010; FU et al., 2004). Automatic

processes to produce ontologies are a constant presence, such as the work of capturing acronyms from Request for Comments (RFCs) corpus, documents produced by the Internet Engineering Task Force (IETF) (BRAGA, 2019a). There are many works related to the Ontologies construction that include methodologies using Protégé, in the most diverse domains and applications [e.g. (REN, 2012; SHEKHAR; K, 2013; CIMIANO; VÖLKER, 2005; BUITELAAR; OLEJNIK; SINTEK, 2004; TAN; LAMBRIX, 2009)]. However, initiatives for the construction of Ontologies that express the knowledge of the domain of Machine Learning have few references on very specific sub-domains (SACHA et al., 2018; DOAN et al., 2004; ICHISE, 2008).

The possibility to complement any ontology with domain ontologies developed by others is proposed by the ontology alignment approach proposed by Ehrig (2007).

Lately, new area is being proposed: ontology learning using techniques of Natural Language Processing, AI and machine learning (BUITELAAR; CIMIANO; MAGNINI, 2005; ASIM et al., 2018; KONYS, 2018) (BIRD; KLEIN; LOPER, 2009) (BHAVSAR; KUMAR; DANGETI, 2017).

3. MLOnto Architecture

Recently, Vaishak Belle and Ioannis Papantonis (BELLE; PAPANTONIS, 2020), divided multi-stakeholders in understanding machine learning models into five groups, when they observed that ML models are increasingly deployed in a wide range of companies: (a) data scientists, (b) business owners, (c) model risk analysts, (d) regulators and (e) consumers, characterizing them in a didactically clear way.

Concerns and experiences like those of these authors, stimulated the creation of the first stage of MLOnto, adding the fact that another interested party, (f) the agents (as shown in Figure 1), must follow the changes and the complexity of the knowledge domain that populates the machine learning discipline.

The first stage of the construction of the MLOnto is manual and we considered the seven main classes presented in Figure 3, whose details are discussed in the subsection below.

These seven classes were considered, for now, sufficient to describe the domain of knowledge that encompasses the discipline of Machine Learning considered a branch of Artificial Intelligence (AI).

3.1. The MLOnto Highest Level Classes

The main characteristics of these seven highest level classes are:

- Algorithms: A set of rules that precisely defines a sequence of operations, which would include all computer programs, including programs that do not perform numeric calculations.
- Applications: the action of putting something into operation.
- **Dependencies**: What are the prerequisites for solving an algorithm or problem and what are the requirements for a tool to be executed.
- **Dictionary**: Acronyms, nicknames and phrases for which an interpretation is required.
- **Frameworks**: A set of tools and associated utilities for solving and implementing algorithms, problems or applications.



Figure 3. The MLO highest level classes

- **Involved**: People or institutions that created and improved the available resources to facilitate the use and application of Machine Learning.
- **MLTypes**: Usual classification of Machine Learning algorithms. Although, more recently, the literature is adopting only three types of Machine Learning (a) Supervised Learning, (b) Unsupervised Learning, (c) Reinforcement Learning we are initially adopting two additional types, (d) AutoML, (e) Semi-Supervised Learning (Figure 4).



Figure 4. Machine Learning Types in MLOnto

4. The implementation structure

The resources available in Protégé will be used extensively, such as the Annotations (Figure 5) that allow indicating the references and clarify or justify non-traditional choices.



Figure 5. Annotation implementation in Protégé

The contents added in Annotations are axioms and therefore react naturally under the mechanisms of inference or search performed by humans or agents. Additionally, properties that can be included in Annotations are dynamic and can be added at the discretion of the ontology administrator. The Figure 6 shows six properties added specifically for MLOnto's interests.



Figure 6. MLO especific Annotations properties

Protégé produces an OWL file that serves agents made by software, also. OWL serves human beings, too, but the Protégé interface is better prepared to facilitate human understanding of ontology. For example, in Figure 7, a view centered on the Algorithms class displays the branches graphically.

In addition, Protégé's features include facilities that allow MLOnto, to debugging allowing corrections and refinements to be made easily with the created ontology.

5. Future Work

The ontology is dynamic, as the knowledge of the Machine Learning domain is constantly renewed. A mechanism is needed to ensure the systematic updating of MLOnto



Figure 7. Partial graphic from the MLOnto with emphasis on the Algorithms class

in view of the Never-Ending Learner paradigm (MITCHELL et al., 2018). Specialized languages in learning that never-ends, have been proposed for applications in large domains, have been studied but can be thought of applying over restricted and very narrow domains, such as in the SciPy tool facility updates (CARLSON et al., 2010) Also, we must consider that the application referred to as Lifelong Machine Learning (LML), considers systems that can learn many tasks from one or more domains over its lifetime (SILVER; YANG; LI, 2013)

It would be appropriate for issues related to ethics in the construction of algorithms and the implementations that associate transparency, law and individual or group values of human society to be incorporated and characterized in MLOnto (ARBIX, 2020).

It is necessary to advance research and efforts in automatic processes for building ontologies from unstructured bases, mainly, encouraging the continuity of similar projects, which already exist (SHAMSFARD; BARFOROUSH, 2004). Additionally, along the same lines and concerns, innovative research must be strongly considered (GUIZZARDI et al., 2015).

There are many inference techniques available and it is necessary to research which ones are most appropriate to the various ontologies and identify the respective applications. Protégé has several inference mechanisms that make testing easy to identify the most appropriate ones.

Ontology Evaluation is the problem of assessing a given ontology from the point of view of a particular criterion of application, tipically in order to determine which of several ontologies would best suit a particular purpose (BRANK; GROBELNIK; MLADENIC, 2005).

In the sequence, it is admitted the proposal to use the Ontop², a popular open-source Virtual Knowledge Graph (VKG) system that can expose heterogeneous data sources

²https://ontop-vkg.org/guide/

as a unified knowledge graph (XIAO et al., 2020) (XIAO et al., 2019). Ontop has a Protégé plug-in facility that will be very useful for SPARQL applications. Also, agents of the A2RD who will benefit from the SKAU environment will have access to SPARQL Ontop facilitating search and reasoning mechanisms (Figure 1).

Ontop, other facilities and resources can and should be checked for ontology alignment or ontology marching, the process that takes two or more ontologies as input and outputs a single ontology, where the similarities are adjusted accordingly. An intuitive view can be seen in Figure 8, whose listed components are detailed at Ehrig (2007, Page 62).



Figure 8. The input, output, and the six main steps involved in alignment process. Adapted from (EHRIG, 2007)

A complementary text with a review of a number of techniques available and associated with ontology alignment can be found at Ardjani, Bouchiha e Malki (2015).

6. Conclusions

An ontology like this, in the domain of Machine Learning is a practically endless task. From a certain point on. the authors' concerns are emphatic in the questions of quality and reliability of the data that will be included. As an example it is worth mentioning the choice to include the techniques and algorithms used in applications under the domain of Internet Infrastructure. Our team considered that this task should be preceded by a literature review (not necessarily a systematic literature review) with which would be obtained the primary works that could confirm the additions to the current ontology.

The example given above, associated with the Internet Infrastructure, was made the choice to include the machine learning applications over the Autonomous Systems sub-domain, name given to the computer networks that represent, together, the Internet (BRAGA, 2019b). Thus, it was a matter of doing a literature review (non-systematic) to survey the primary studies that could produce the data and qualify them.

7. Data availability

The latest version of the MLOnto database generated by the current study is publicly available on Open Science Foundation³ (OSF) (BRAGA; REGATEIRO; DIAS, 2020). Some metrics from MLOnto in Protégé is available in Figure 9. Content license: Creative Commons 4.0 BY.

8. Conflicts of interest

The authors declare that there are no conflicting interests.

```
<sup>3</sup><https://osf.io/chu5q/>
```

Ontology metrics:	2 🛛 🗖 🗖 🖻
Metrics	
Axiom	1877
Logical axiom count	1157
Declaration axioms count	496
Class count	433
Object property count	56
Data property count	1
Individual count	1
Annotation Property count	11

Figure 9. Some metrics from MLOnto in Protégé

References

ANTONIOU, G.; HARMELEN, F. V. Web ontology language: Owl. In: *Handbook on ontologies*. [S.l.]: Springer, 2004. p. 67–92.

ARBIX, G. A trAnspArênciA no centro dA construção de umA iA éticA. *Novos estudos CEBRAP*, SciELO Brasil, v. 39, n. 2, p. 395–413, 2020.

ARDJANI, F.; BOUCHIHA, D.; MALKI, M. Ontology-alignment techniques: Survey and analysis. *International Journal of Modern Education & Computer Science*, v. 7, n. 11, 2015.

ASIM, M. N.; WASIM, M.; KHAN, M. U. G.; MAHMOOD, W.; ABBASI, H. M. A survey of ontology learning techniques and applications. *Database*, Narnia, v. 2018, 2018.

BELLE, V.; PAPANTONIS, I. Principles and practice of explainable machine learning. *arXiv preprint arXiv:2009.11698*, 2020.

BHAVSAR, K.; KUMAR, N.; DANGETI, P. Natural language processing with python cookbook: Over 60 recipes to implement text analytics solutions using deep learning principles. Packt Publishing, 2017.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python*. [S.l.]: " O'Reilly Media, Inc.", 2009.

BORST, W. N. *Construction of engineering ontologies for knowledge sharing and reuse*. Tese (Doutorado) — University of Twente, 1997.

BRAGA, J. Ambiente para Aquisição de Conhecimento por Agentes em Domínios Restritos na Infraestrutura da Internet. Tese (Doutorado) — Instituto Superior Técnico & Universidade Presbiteriana Mackenzie, 2019. DOI: 10.31237/osf.io/nzmtf, Availaible in <https://thesiscommons.org/nzmtf/>.

BRAGA, J. Environment for Knowledge Acquisition by Agents in Internet InfrastructureRestricted Domains. Tese (Doutorado) — Instituto Superior Tecnico & Universidade Presbiteriana Mackenzie, 2019. DOI: 10.31237/osf.io/83ztf. English version translated by author. Available in https://thesiscommons.org/83ztf/.

BRAGA, J.; DIAS, J. L. R.; REGATEIRO, F. A machine learing ontology. *Frenxiv*, Oct 2020. Preprint. DOI: 10.31226/osf.io/rc954. Available at https://frenxiv.org/rc954/>.

BRAGA, J.; NOBRE, J. C.; GRANVILLE, L. Z.; SANTOS, M. Como Protocolos Inovadores são Criados e Adotados em Escala Mundial: Uma visão sobre o Internet Engineering Task Force (IETF) e a Infraestrutura da Internet. In: WEBER, T. S.; MARTINS, C. A. (Ed.). *Jornadas de Atualização em Informática 2020*. Cuiabá, MT Brazil: Sociedade Brasileira de Computação, 2020. p. 45. ISBN 978-65-87003-28-3. Available in: https://doi.org/10.5753/sbc.5728.3.2.

BRAGA, J.; REGATEIRO, F.; DIAS, J. *Machine Learning Ontology (MLOnto) Repository*. 2020. Project Repository: OSF. DOI: 10.17605/OSF.IO/CHU5Q. Disponível em: <osf.io/chu5q>.

BRAGA, J.; SILVA, J. N.; ENDO, P.; OMAR, N. Structure for knowledge acquisition, use, learning and collaboration inter agents over internet infrastructure domains:
Proceedings of the 2019 computing conference. In: ARAI, K.; BHATIA, R.; KAPOOR, S. (Ed.). *Intelligent Computing*. [S.1.]: Springer International Publishing, 2019. v. 1, p. 527–547. Doi: 10.1007/978-3-030-22871-2, ISBN: 978-3-030-22871-2.

BRANK, J.; GROBELNIK, M.; MLADENIC, D. A survey of ontology evaluation techniques. In: *Proceedings of the conference on data mining and data warehouses* (*SiKDD 2005*). Ljubljana, Slovenia: SiKDD, 2005. p. 166–170.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, IOS Press, Amsterdam, v. 123, p. 3–12, 2005.

BUITELAAR, P.; OLEJNIK, D.; SINTEK, M. A protégé plug-in for ontology extraction from text based on linguistic analysis. In: SPRINGER. *European Semantic Web Symposium*. [S.1.], 2004. p. 31–44.

CARLSON, A.; BETTERIDGE, J.; KISIEL, B.; SETTLES, B.; HRUSCHKA, E. R.; MITCHELL, T. M. Toward an architecture for never-ending language learning. In: *Twenty-Fourth AAAI Conference on Artificial Intelligence*. Atlanta, US: AAAI Publications, 2010. p. 8.

CIMIANO, P.; VÖLKER, J. A framework for ontology learning and data-driven change discovery. In: SPRINGER. *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*. [S.1.], 2005. p. 227–238.

DOAN, A.; MADHAVAN, J.; DOMINGOS, P.; HALEVY, A. Ontology matching: A machine learning approach. In: *Handbook on ontologies*. [S.l.]: Springer, 2004. p. 385–403.

EHRIG, M. Ontology Alignment: Bridging the Semantic Gap. 1. ed. Germany: Springer, 2007.

FU, H.; CHI, Z.; FENG, D.; SONG, J. Machine learning techniques for ontology-based leaf classification. In: IEEE. *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.* [S.1.], 2004. v. 1, p. 681–686.

GOLBREICH, C.; DAMERON, O.; GIBAUD, B.; BURGUN, A. Web ontology language requirements wrt expressiveness of taxonomy and axioms in medicine. In: SPRINGER. *International Semantic Web Conference*. [S.1.], 2003. p. 180–194. GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, v. 5, p. 199–220, 1993.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: *Handbook on ontologies*. [S.l.]: Springer, 2009. p. 1–17.

GUIZZARDI, G.; WAGNER, G.; ALMEIDA, J. P. A.; GUIZZARDI, R. S. Towards ontological foundations for conceptual modeling: The unified foundational ontology (ufo) story. *Applied ontology*, IOS Press, v. 10, n. 3-4, p. 259–271, 2015.

HITZLER, P.; KRöTZSCH, M.; PARSIA, B.; PATEL-SCHNEIDER, P. F.; RUDOLPH, S. OWL 2 Web Ontology Language Primer. *W3C Recommendation*, v. 27, n. 1, p. 1–123, 2009. Disponível em: ">https://www.w3.org/TR/owl2-primer/>.

HORRIDGE, M. A Practical Guide To Building OWL Ontologies Using Prot´eg´e 4 and CO-ODE Tools Edition 1.3. [S.1.], 2011. 108 p. Disponível em: https://www.researchgate.net/publication/230585369_A_Practical_Guide_To_Building_OWL_Ontologies_Using_The_Prot%27eg%27e-OWL_Plugin_and_CO-ODE_Tools.

ICHISE, R. Machine learning approach for ontology mapping using multiple concept similarity measures. In: IEEE. *Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008)*. [S.1.], 2008. p. 340–346.

KALIBATIENE, D.; VASILECAS, O. Survey on ontology languages. In: SPRINGER. *International Conference on Business Informatics Research*. [S.l.], 2011. p. 124–141.

KONYS, A. Knowledge systematization for ontology learning methods. *Procedia computer science*, Elsevier, v. 126, p. 2194–2207, 2018.

MITCHELL, T.; COHEN, W.; HRUSCHKA, E.; TALUKDAR, P.; YANG, B.; BETTERIDGE, J.; CARLSON, A.; DALVI, B.; GARDNER, M.; KISIEL, B. et al. Never-ending learning. *Communications of the ACM*, ACM, v. 61, n. 5, p. 103–115, 2018.

MUSEN, M. A. The protégé project: a look back and a look forward. *AI Matters*, v. 1, n. 4, p. 4–12, 2015. Disponível em: https://doi.org/10.1145/2757001.2757003>.

PEASE, A. Ontology. [S.1.]: Articulate Software Press, 2011. ISBN: 1889455105.

POON, H.; DOMINGOS, P. Unsupervised ontology induction from text. In: *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*. [S.1.: s.n.], 2010. p. 296–305.

REN, F. A demo for constructing domain ontology from academic papers. In: *Proceedings of COLING 2012: Demonstration Papers*. [S.l.: s.n.], 2012. p. 369–376.

SACHA, D.; KRAUS, M.; KEIM, D. A.; CHEN, M. Vis4ml: An ontology for visual analytics assisted machine learning. *IEEE transactions on visualization and computer graphics*, IEEE, v. 25, n. 1, p. 385–395, 2018.

SHAMSFARD, M.; BARFOROUSH, A. A. Learning ontologies from natural language texts. *International journal of human-computer studies*, Elsevier, v. 60, n. 1, p. 17–63, 2004.

SHEKHAR, M.; K, S. R. Semantic Web Search based on Ontology Modeling using Protege Reasoner. 2013.

SILVER, D. L.; YANG, Q.; LI, L. Lifelong machine learning systems: Beyond learning algorithms. In: 2013 AAAI spring symposium series. [S.l.: s.n.], 2013.

SLIMANI, T. Ontology development: A comparing study on tools, languages and formalisms. *Indian Journal of Science and Technology*, v. 8, n. 24, p. 1–12, 2015.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. *Data & knowledge engineering*, Elsevier, v. 25, n. 1-2, p. 161–197, 1998.

TAN, H.; LAMBRIX, P. Selecting an ontology for biomedical text mining. In: *Proceedings of the BioNLP 2009 Workshop*. [S.l.: s.n.], 2009. p. 55–62.

TUDORACHE, T.; NYULAS, C.; NOY, N. F.; MUSEN, M. A. Webprotégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic web*, IOS Press, v. 4, n. 1, p. 89–99, 2013.

XIAO, G.; DING, L.; COGREL, B.; CALVANESE, D. Virtual knowledge graphs: An overview of systems and use cases. *Data Intelligence*, MIT Press, v. 1, n. 3, p. 201–223, 2019.

XIAO, G.; LANTI, D.; KONTCHAKOV, R.; KOMLA-EBRI, S.; GÜZEL-KALAYCI, E.; DING, L.; CORMAN, J.; COGREL, B.; CALVANESE, D.; BOTOEVA, E. The virtual knowledge graph system ontop. In: SPRINGER. *International Semantic Web Conference*. [S.1.], 2020. p. 259–277.