12-2-2023

An Application Ontology for Reproducibility of Machine Learning Solutions

Madhushi Bandara
University of Technology Sydney, Australia, madhushi.bandara@uts.edu.au

Yuchao Jiang
University of New South Wales, Australia, yuchao.jiang@unsw.edu.eu

Asif Gill
University of Technology Sydney, Australia, asif.gill@uts.edu.au

Fethi A. Rabhi
University of New South Wales, Australia, f.rabhi@unsw.edu.au

Ghassan Beydon
University of Technology Sydney, Ghassan.Beydoun@uts.edu.au

An Application Ontology for Reproducibility of Machine Learning Solutions

Full research paper

Madhushi Bandara
School of Computer Science
University of Technology Sydney
Sydney, Australia
Email: madhushi.bandara@uts.edu.au

Yuchao Jiang
School of Computer Science & Engineering
University of New South Wales
Sydney, Australia
Email: yuchao.jiang@unsw.edu.au

Asif Gill
School of Computer Science
University of Technology Sydney
Sydney, Australia
Email: asif.gill@uts.edu.au

Fethi A. Rabhi
School of Computer Science & Engineering
University of New South Wales
Sydney, Australia
Email: f.rabhi@unsw.edu.au

Ghassan Beydoun
School of Computer Science
University of Technology Sydney
Sydney, Australia
Email: ghassan.beydoun@uts.edu.au

Abstract

With Artificial Intelligence and Machine Learning (ML) on the rise, organisations of different scales and nature are looking to utilise ML systems to support their day-to-day operations. Many enterprises find it difficult to adapt existing ML solutions to their organisations without huge investments in solution understanding, customisation, infrastructure enablement and workforce training. Some organisations utilise external service providers to provision their standard analytics services, and this often leads to solutions that either do not fit well with their organisation goals or may lead to the loss of expert knowledge behind the establishment of the AI system. This paper aims to address some of these challenges by proposing an ontology for ensuring the reproducibility of ML models in research as well as their integration within application environments. Our work will ensure that the knowledge about a developed ML system or process is accumulated and recorded within an organisation and can be used in the future, either by new employees or other teams within the organisation. This approach can also be utilised by researchers and developers of ML systems to record and publish metadata of their studies, ensuring that future researchers can reuse their work with minimal effort.

Keywords: Machine Learning, reproducibility, ontology, knowledge graph

1 Introduction

Organisations of all scales seek to utilise machine learning to analyse their operations and support their everyday decision-making. However, the research and development of machine learning techniques and platforms are fast paced, requiring extensive technical expertise and domain expertise to deploy. A large number of new publications are constantly proposing new and improved machine learning models. There are no systems yet in place to record relationships between data, analytics steps nor the results from the computational experiments and non-computational aspects such as interpretation (Samuel and König-Ries 2022a).

ML model solutions in an organisation are usually developed by an outsourced expert team. These models are rarely utilised beyond the scope of research or specific project, reused, or translated into other settings. Any knowledge accumulated by engineers is lost over time. A recent survey found that researchers outside the machine learning core community find it quite difficult to keep up with the litany of terminology, techniques, and metrics developed. The increasing diversity and application of machine learning approaches make it also increasingly difficult to discern their intent and provenance (Leipzig et al. 2021). Commercial analytic platforms that incorporate latest ML models exist but utilising or customising them for a given business context can be very expensive. Clearly, organisations can benefit from metadata that describes those models' operations, assumptions, and connections to the IT infrastructure, so that they can be maintained and reused without necessarily involving the analysts who created them in the first place. Systems that are in place to ensure research reproducibility need to capture and report extensive metadata about the data analytics experiments, including domain and scientific concepts, and the underlying system environment (Leipzig et al. 2021). This can be of many forms, from the generalisation of underlying scientific findings and exact recreation of an experiment to open sharing of the analysis for future use (Leipzig et al. 2021). Indeed, premier machine-learning conferences and journals have identified this issue and often request authors to submit a reproducibility checklist with their work. There is also an emerging body of research looking into ML reproducibility. They all rely on documentation and checklists provided individual project teams. They often lack a standard format or vocabulary and are not easy to integrate into IT systems for automated recording or querying. Moreover, when completing a checklist, there can be various interpretations. Thus, there is a clear need for formal vocabularies and standards to represent and measure reproducibility in ML systems by capturing relevant metadata.

In this paper, we propose the use of a semantic model, or ontology, to capture extensive metadata about ML models and processes in a semi-structured format to ensure ML output understandability and reproducibility. The utility of ontologies as formal information models to capture complex information has been demonstrated in various research domains, including the biomedical domain, where ontologies such as SNOMED and Gene ontology are utilised to establish recording standards, facilitate information sharing, and design digital systems to search and infer knowledge (Bandara et al. 2019).An ontology can provide a formal yet flexible schema to capture complex and interrelated metadata. ML model details can then be systematically recorded, and users can easily search and access the information to learn about ML models and evaluate their reproducibility. Hence, we have developed the ML-Reproduce Ontology following the guidelines of Design Science Research Methodology (DSRM) (Gregor and Hevner 2013). Our work advances the understanding of the ML reproducibility problem by providing an information model to conceptualise factors that facilitate ML reproducibility and their integration based on multiple insights. From a practical perspective, the proposed ML-Reproduce ontology brings value to analytics systems in the following ways:

- It captures and stores the experience and knowledge accumulated by ML engineers during the development of ML models in a knowledge base using semantic modelling principles.
- The knowledge base maintains relationships amongst prediction models, data elements and variables, their corresponding data sources, their performance and documentation. It enables analysts to record analytics models and makes them readily available and modifiable for future use and adaptation by other stakeholders.
- This knowledge base can be used as a guideline for creating an ML reproducibility checklist or a metric for reproducibility.
- ML-Reproduce can support ML communities in better organisation, communication, and reusability of knowledge about analytics models. Arguably, no two ML models or processes are identical. Each tends to have unique analytics requirements. ML-Reproduce provides the ability to describe their target communities and to transfer and extend models between different applications catering for unique contexts as they arise.

We evaluated the functionality and utility of the ML-reproduce through a prototype implementation in the Stardog knowledge graph platform, for the use case of Reproducibility in deep learning based medical image segmentation.

The remainder of this paper is organised as follows: Section 2 introduces related work. Section 3 presents our solution. Section 4 presents use case and scenarios to evaluate the utility of our platform. Finally, Section 5 concludes with a summary and future work.

2 Related Work

ML solution development consists of many aspects, from the propagation of raw data collected from different sources, intermediate data structures, and computational hardware to open code and statistical analysis, and finally, publication (Leipzig et al. 2021). The conventional way of recording ML experiments and pipeline development is through documentation within the code or using a tool such as Jupyter Notebook. Furthermore, user guides and instructions (e.g., Readme file) are shared with users. Such recording systems lack the integration of information around data, steps, and results from the complete processes of the experiment (Samuel and König-Ries 2022a).

Quantifying the reproducibility of scientific research is not straightforward. Gundersen et al. (Gundersen and Kjensmo 2018) has proposed six metrics measuring three different degrees (Method, Data, and Experiment) of reproducibility in empirical AI research. Neurips and other highly reputed AI/ML conferences measure reproducibility through a checklist (e.g., guideline for 2022 Neurips[1]). Raff et al. (Raff 2019) tried to reproduce ML algorithms independently from papers without using published code and computed the association of reproducibility to different features. Machicao et al. (Machicao et al. 2022) has also developed a comprehensive checklist as a mitigation strategy to ensure reproducibility in deep learning models for remote sensing. Renard et al. (Renard et al. 2020) also looked into the issue of Reproducibility, by identifying sources of variability in deep learning models for medical image segmentation and call for better description frameworks and analysis techniques to understand variability and reproduciblity.

Leipzig et al., (Leipzig et al. 2021) reviewed a wide range of metadata standards for reproducible computational research across an “analytic stack” consisting of input data, tools, reports, pipelines, and publications. Their findings advocate for traditional checklists to be integrated with metadata schemas to make them machine-readable and auto-generated, and publications and notebooks to be annotated with inline semantic data. A semantic web technology stack, with an ontology as its basis, has the potential to realise these features.

Ontologies provide a systematic explanation of things and have received attention in IS discipline as means for domain conceptualisation, standardisation, interoperability, and bridging knowledge among stakeholders (Kishore and Sharman 2004). Ontologies provide structure and codify knowledge about concepts, relationships, and axioms/constraints in a computational format for manipulation by stakeholders and systems. Ontologies are supported by many languages and technologies such as semantic web standards and are often utilised to manage knowledge intensive processes (Bandara et al. 2018).

Few ontologies in the literature attempt to address the issue of reproducibility in research settings, particularly in clinical and healthcare domains. One example is the ProvCaRe model (Liu et al. 2021), a repository of provenance metadata extracted from published biomedical research studies. It has been utilised in NLP extraction, yet it is a simple high-level model that cannot be utilised to capture complex metadata required for ML reproducibility. Another example is the REPRODUCE-ME ontology (Samuel and König-Ries 2022a; Samuel and König-Ries 2022b) designed for scientific experiments in the biomedical domain, which provides a mature set of tools, including an extension for Jupyter Notebooks (Samuel and König-Ries 2018) to support computational reproducibility. It covers high-level concepts such as data, agents, activities, plans, steps, variables, and instruments but does not contain granular classes and attributes necessary to model ML experiment reproducibility. None of the identified studies specialise at ML reproducibility, nor are agonistic to one application domain.

Research Variable Ontology (RVO) is another approach that aims to capture analytics-related knowledge, particularly how datasets and variables are associated within a machine learning model (Bandara et al. 2019). RVO has been used to support analytics platform design and visualisation, with

[1] https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

RVO-based knowledge-base supporting requirements management and end-user decision making (Bandara et al. 2023, Rabhi et al. 2021).

This paper plans to extend RVO and propose a more comprehensive meta-data model for ML reproducibility that can be integrated with existing digital platforms and applied into a wide range of ML problems.

3 Research Method

This paper develops and tests an ontology through which ML engineers and users can communicate via a shared language and unify the implementation of ML systems catering to ML reproducibility. We followed the well-known guidelines from the Design Science Research Methodology (DSRM) (Gregor and Hevner 2013) for developing and evaluating a practical artefact to address the research problem in hand (Peffers et al. 2018).

DSRM defines five phases of (i) Identification of the problem and motivation, (ii) definition of the objectives (iii) design and development, (iv) demonstration, (v) evaluation and (vi) communication.

The first phase and second phase of the method are associated with the identification of a justifiable research problem, followed by defining an objective solution to solve the research problem. As mentioned earlier, the study motivation is the lack of a systematic approach to capture and disseminate knowledge on ML reproducibility. Our objective is to propose an ontology to capture extensive metadata about ML models and processes to solve this challenge.

In the design phase, we adopted ontologies as the theoretical lens, and followed NeOn approach (Suárez-Figueroa et al. 2015) that suggests reusing non-ontological resources and ontology design patterns in an ontology design endeavour. It extends existing approaches by providing explicit guidelines and definitions to create networked collaborative ontologies. NeON approach was selected because of the simplicity of its steps in ontology design, being a scenario-based approach, and due to the availability of its supporting documentation. The process included formulating Competency Questions (CQs) and identifying domain variables and their relationships to represent CQs. Competency Questions determine the scope and set of statements that the ontology should be able to answer (Gruninger 1995). CQs, which can be described in the natural language, help specify the ontology's boundary and granularity, and identify relevant domain concepts and properties and their relationships (Ren et al. 2014). They can then be coded into formal RDF/OWL modelling language.

In the demonstration phase, ML-reproduce ontology was implemented as a prototype in the StarDog enterprise knowledge graph platform[2], and populated with instance data based on the deep learning based medical image segmentation case study discussed in section 4. We evaluated the efficacy of the ontology to capture the reproducibility attributes of an ML model, by evaluating how the prototype implementation can be utilised to answer competency questions. The lessons learned from this project are communicated to academia and practice in the communication phase.

4 Proposed Solution- Ontology Design and Development

Following the research method outlined above, we gathered prerequisites and requirements to design and develop the ontology through a literature review. Based on our literature review, we identified three key studies that have catalogued the checklists and dimensions of reproducibility in ML technologies (Machicao et al. 2022 Gundersen and Kjensmo 2018; Samuel and König-Ries 2022b). Based on their work, we compiled the ontology requirements and 67 competency questions organised under six different aspects: problem context, model architecture, process and provenance, dataset, results and claims, and program source code. Table 1 provides few randomly selected sample competency questions; please refer to the supplementary resources[3] for the complete list of questions.

Once the concepts are identified, we based our ML-Reproduce ontology design on RVO- The Research Variable Ontology[4]. RVO contained primary concepts necessary to design a model, such as ML Model and Dataset. The resulting model is visualised in Figure 1. Concepts from external ontologies also integrated, relying on their specificity for respective domains, and to ensure ML-Reproduce can be

[2] https://www.stardog.com
[3] https://docs.google.com/document/d/1fTEMwDN7zxnLf2Y-o8oCOFoYrYjEmUNE/
[4] http://adage2.cse.unsw.edu.au/rvo/

integrated with other ontology-based systems. For example, FOAF ontology[5] was used to describe Person concept and FaBIO[6] was used to describe research papers associated with models.

CQ ID	CQ	Related aspect
CQ5	What are the input and output variables of an ML model?	Model Architecture
CQ13	Which parameter tuning techniques were used?	Model Architecture
CQ32	Which are the datasets that were used in an ML experiment?	Dataset
CQ35	What data were excluded, and for what reason?	Dataset

Table 1. Sample Competency Questions

ML-Reproduce was implemented using OWL (Web Ontology Language) (Bechhofer et al. 2004), the common language to create ontologies including classes/concepts, subclasses, properties, and associated relationships of a domain of interest. We used Protégé[7] open-source ontology editor tool that supports RDF/OWL standards.

Once the model was created, it was imported into Stardog knowledge graph modelling platform. Figure 1 shows the ontology visualisation through Stardog Explorer interface. To demonstrate how ML-Reproduce ontology can be used to represent metadata associated with ML reproducibility, we populated the Stardog platform with sample instance data extracted from 23 machine learning solutions reviewed by Renard et al. (Renard et al. 2020). Details of the data extraction and implementation are presented as a use case in Section 5. Example queries of this use case, running in Stardog platform using SPARQL are shown in Figures 2 and 3. This implementation and evaluation helped us identify limitations and design issues with ML-Reproduce and improve it in an iterative manner.

5 Evaluation

To evaluate the functionality and utility of the ML-Reproduce ontology, we implemented a literature-based use case in the Stardog knowledge graph platform. We extracted meta-data on 23 machine learning solutions identified by Renard et al. (Renard et al. 2020) and used Stardog platform to encode them as instances of ML-Reproduce. The 23 machine learning solutions they identified were proposing deep learning solutions for medical image segmentation. We decided to use the same set of studies, as that gives us the ability to compare the completeness of our ML-reproduce model with the findings reported by Renard et al. Our observation was that we were able to successfully capture all the attributes and dimensions of the 23 ML models reported in their work.

For example, openly shared code is of high significant in research reproducibility. So, in Figure 2, we have written a SPARQL query to identify ML models in the knowledge graph that have their source code published as open source and to retrieve their model type and description. The output for the query shows three ML models, all three are of CNN model type.

If users want to learn further about one particular model, they can write a detailed SPARQL query indicating the properties and values they are interested in. For example, in Figure 3, the query retrieves meta-data about the model developed by Kamnitsas et al. (Kamnitsas et al. 2017), indicating that its implementation framework is Theano, its cross-validation strategy is k-fold, and optimisation is conducted by RMS-prop.

In the next step of the evaluation, we investigated the ability of ML-Reproduce to answer all the competency questions identified as functional requirements in the ontology design stage. The full list of 67 competency questions is available online[3]. We looked into the ML-Reproduce implementation to assess what questions can be answered through a SPARQL query to the current model. Our findings are recorded under Evaluation Outcome column of the Competency Questions.

We identified that 45 CQs were successfully addressed by ML-Reproduce, and 2 partially addressed. We observe that 11 out of the successful CQs were inherent by design, as ML-Reproduce is modelled

[5] https://en.wikipedia.org/wiki/FOAF
[6] https://sparontologies.github.io/fabio/current/fabio.html
[7] https://protege.stanford.edu

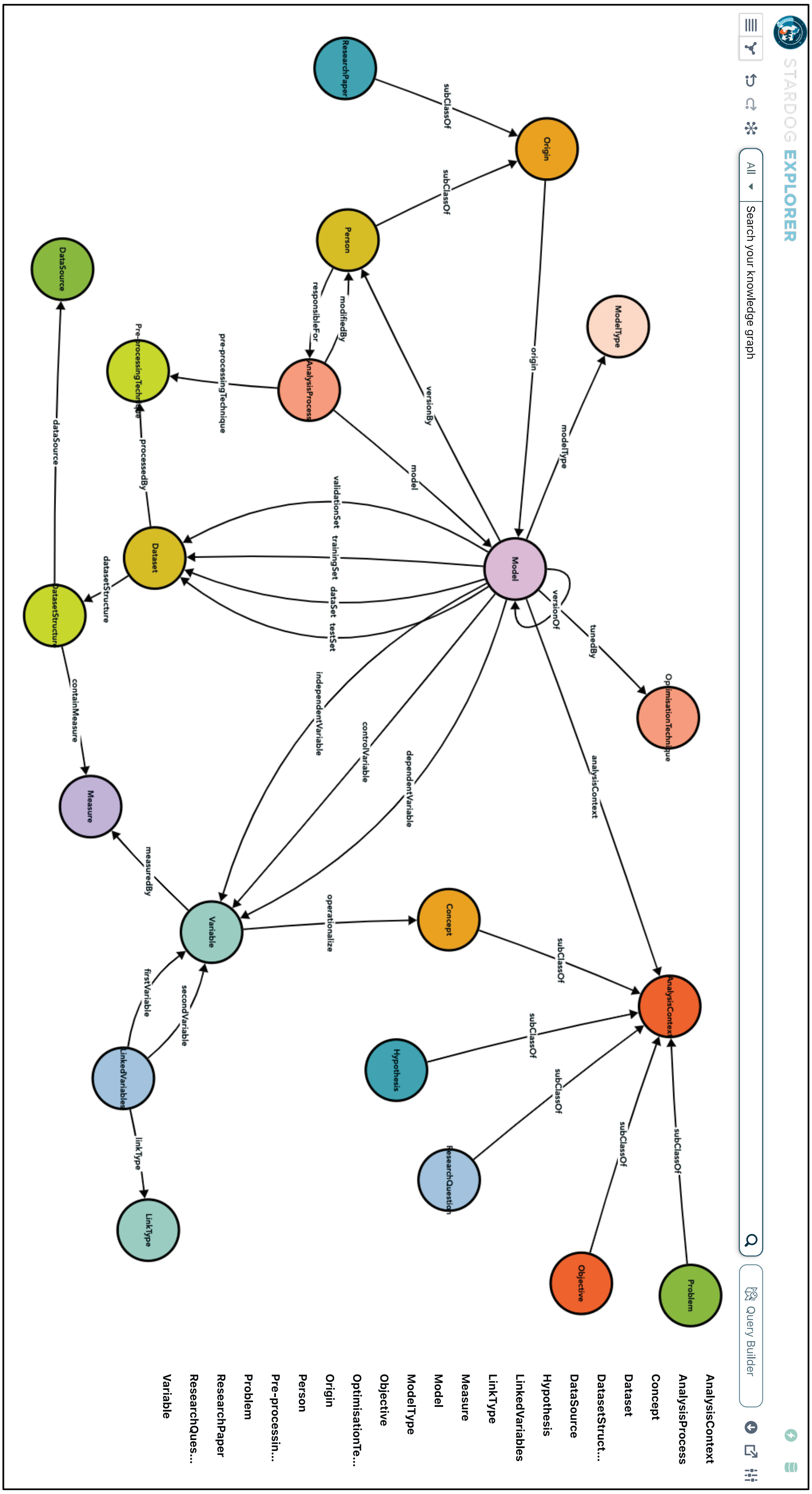

Figure 1. Main classes and relationships of the ML-Reproduce ontology.

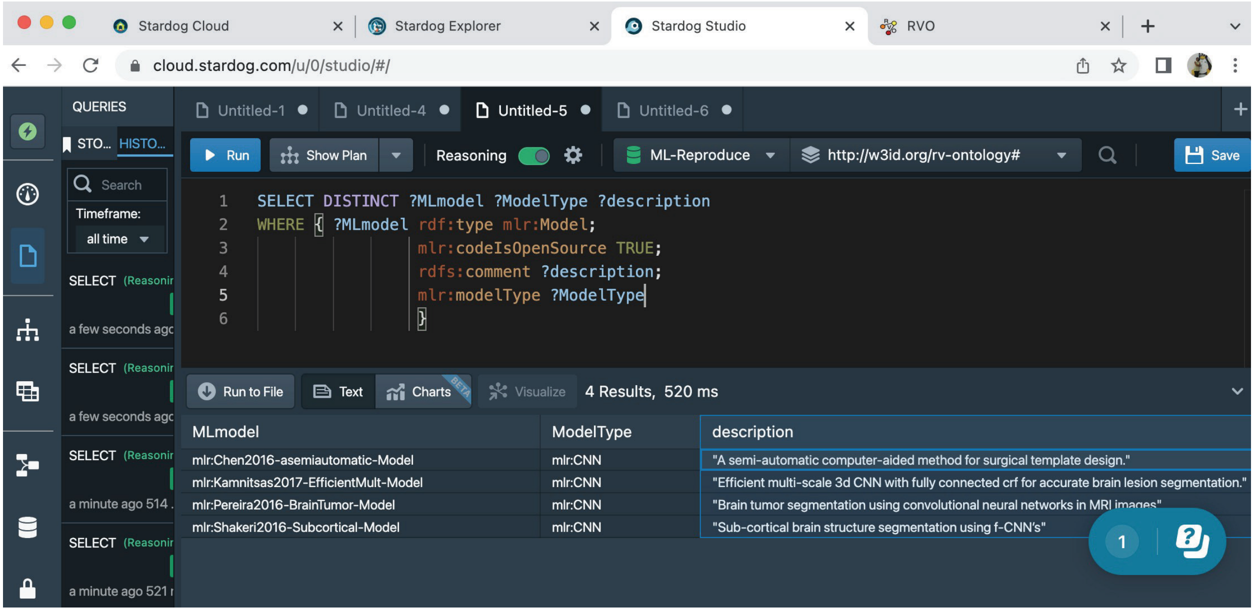

Figure 2. SPARQL query to identify all ML models with open source code.

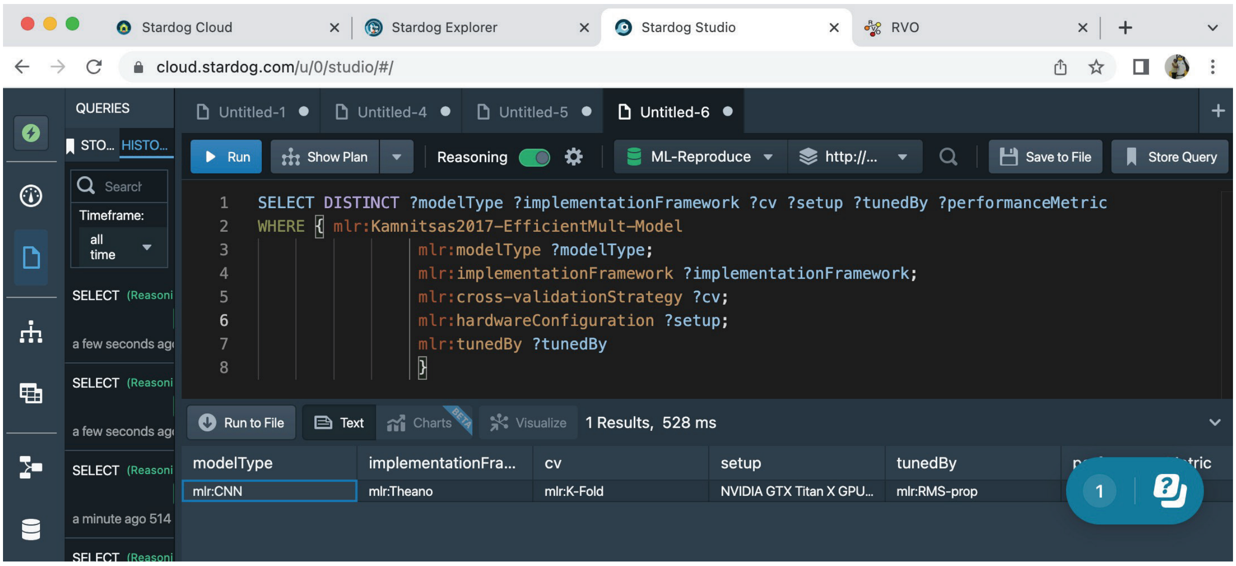

Figure 3. SPARQL query to get details about a model of interest.

following semantic web standards, and extending RVO. Specially, all the FAIR principles (guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets[8]) are automatically adhered to, when the meta-data is recorded following semantic web standards such as using URIs (Uniform Resource Identifiers) and class structure. 20 of the competency questions were not answered through ML-Reproduce, and all of them, except one are associated with the program source code and execution details. This is a current limitation of ML-Reproduce that we plan to address in the next iteration of development by integrating a module for automated meta-data extraction from the code or notebook itself.

[8] https://www.go-fair.org/fair-principles/

Also, the CQ 47 about Results and Claims- 'Is the proof of claim reported clearly?' is not addressed through ML-Reproduce, as this requires expert judgement. When comparing with check-list based methods such as Machicao et al. (Machicao et al. 2022), we conclude that ML-Reproduce based systems can contribute to provide context, structure, and automation in recording meta-data associated with reproduction of ML models, as well as when evaluating Reproducibility of existing ML models. Where the checklist-based methods can answer questions like "Does the paper detail the infrastructure adequately?", an ontology, can easily record, organise and query for more information such as "What are the details of the implementation infrastructure?" or "What are the models that use Theano implementation infrastructure?".

6 Conclusion

This paper proposes the ML-Reproduce ontology, a semantic information model to capture extensive meta-data necessary to aid the reproducibility of ML solutions. Our work helps with capturing and reusing knowledge in developing ML systems in an organisation, for example, by new employees or other teams within the organisation. ML-Reproduce can also be utilised by researchers and developers of ML systems to annotate and publish relevant metadata with their studies, ensuring too that future researchers can reuse their work with minimal effort.

We illustrate the functional utility of the ML-Reproduce ontology through a prototype implementation that records ML solutions in a deep learning-based image segmentation application. But a limitation of this study is that competency questions are yet to be validated for broad ML knowledge reuse. Even though all three studies we utilised to generate competency questions are founded on extensive literature reviews, ML-Reproduce still needs to be independently assessed with empirical inquiry involving ML practitioners and researchers in practice. The evaluation conducted in deep learning-based image segmentation models is very promising, but there is need for further evaluation in different model types and analytic applications. Furthermore, a key limitation of current ML-Reproduce model identified through evaluation is its inability to capture program source code and execution details. We plan to address this by extending the model and implementing automated extraction of execution records.

We envision that ML-Reproduce will be able to address the reproducibility research gap identified by Renard et al. (Renard et al. 2020) by providing a vocabulary to better understand and compare different ML models. This will enable the development of tools that can analyse and quantify their reproducibility. As future work, we plan to develop a prototype platform to showcase how ML-Reproduce can be integrated into systems and extend the ontology to ensure explainability of ML solutions. Furthermore, the ontology-based design needs expert intervention for aspects such as CQ47. We believe that with more ML model meta-data accumulated in the knowledge base over time, an ontology-based inference methods can be utilised to accurately assess complex aspects such as clarity of the proof of claims.

7 References

Bandara, M., Behnaz, A., and Rabhi, F. A. 2019. "RVO-the research variable ontology," *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16, Springer*, pp. 412–426

Bandara, M., Rabhi, F. A. and Bano, M. 2023. "A knowledge-driven approach for designing data analytics platforms." *Requirements Engineering* 28.2: 195-212.

Bandara, M., Rabhi, F. A., and Meymandpour, R. 2018. "Semantic model-based approach for knowledge intensive processes." *Software Process Improvement and Capability Determination: 18th International Conference, SPICE 2018, Thessaloniki, Greece, October 9–10, 2018, Proceedings 18*. Springer International Publishing.

Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., Stein, L. A., et al. 2004. "OWL web ontology language reference," W3C recommendation (10:2), pp. 1–53.

Gregor, S. and Hevner, A. R. 2013. "Positioning and presenting design science research for maximum impact," MIS quarterly (), pp. 337–355.

Gruninger, M. 1995. "Methodology for the design and evaluation of ontologies," in Proc. IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing,

Gundersen, O. E. and Kjensmo, S. 2018. "State of the art: Reproducibility in artificial intelligence," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32. 1.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. 2017. “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” Medical image analysis (36), pp. 61–78.

Kishore, R. and Sharman, R. 2004. “Computational ontologies and information systems I: foundations,” Communications of the Association for Information Systems (14:1), p. 8.

Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., and Greenberg, J. 2021. “The role of metadata in reproducible computational research,” Patterns (2:9), p. 100322.

Liu, C., Kim, M., Rueschman, M., and Sahoo, S. S. 2021. “ProvCaRe: A Large-Scale Semantic Provenance Resource for Scientific Reproducibility,” Provenance in Data Science, Springer, pp. 59–73.

Machicao, J., Ben Abbes, A., Meneguzzi, L., Corrêa, P., Specht, A., David, R., Subsol, G., Vellenich, D., Devillers, R., Stall, S., et al. 2022. “Mitigation strategies to improve reproducibility of poverty estimations from remote sensing images using deep learning,” Earth and Space Science (9:8), e2022EA002379.

Peffers, K., Tuunanen, T., and Niehaves, B. 2018. Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Rabhi, F. A., Bandara, M., Lu, K., and Dewan, S. 2021. “Design of an innovative IT platform for analytics knowledge management,” Future Generation Computer Systems (116), pp. 209–219.

Raff, E. (2019). “A step toward quantifying independently reproducible machine learning research,” Advances in Neural Information Processing Systems (32).

Ren, Y., Parvizi, A., Mellish, C., Pan, J. Z., Van Deemter, K., and Stevens, R. 2014. “Towards competency question-driven ontology authoring,” The Semantic Web: Trends and Challenges: 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings 11, Springer, pp. 752–767.

Renard, F., Guedria, S., Palma, N. D., and Vuillerme, N. 2020. “Variability and reproducibility in deep learning for medical image segmentation,” Scientific Reports (10:1), pp. 1–16.

Samuel, S. and König-Ries, B. 2018. “ProvBook: Provenance-based Semantic Enrichment of Interactive Notebooks for Reproducibility.” ISWC (P&D/Industry/BlueSky),

Samuel, S. and König-Ries, B. 2022a. “A collaborative semantic-based provenance management platform for reproducibility,” PeerJ Computer Science (8), e921.

Samuel, S. and König-Ries, B. 2022b. “End-to-End provenance representation for the understandability and reproducibility of scientific experiments using a semantic approach,” Journal of biomedical semantics (13:1), pp. 1–17.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernandez-Lopez, M. 2015) “The NeOn Methodology framework: A scenario-based methodology for ontology development,” Applied ontology (10:2), pp. 107–145.

Copyright