A Semantic Framework for Evaluating Post-hoc Explanations in Link Prediction

Laura Balbi[1,†], Felix Bindt[2,†], Katja Breitenfelder[3,†], Riccardo Campi[4,†], Jitse De Smet[5,†] and Claudia d'Amato[6,†]

[1]LASIGE, Faculty of Sciences, University of Lisbon
[2]Wageningen University & Research
[3]Fraunhofer Institute for Building Physics
[4]Data Science Lab, Polytechnic University of Milan
[5]Ghent University
[6]Department of Computer Science, University of Bari

Abstract

Knowledge Graphs (KGs) are often noisy or incomplete, and Link Prediction (LP) methods, especially those based on black-box KG-Embeddings, are employed to predict missing facts. Pushed by the need for trust in inferred facts, many methods for LP explanation (LP-X) have been created. However, comparing them is still an open issue due to multiple existing protocols. To address this gap, we envision the design of an automated and unified evaluation framework for post-hoc LP-Xs that allows for a systematic and operationalized computation and comparison of LP explanations. To offer a pragmatic view of our proposition, we extend the Explanation Ontology (EO) by enriching it with evaluation-specific constructs, thus providing a shared semantic model (i.e., a structured knowledge representation such as an ontology) that unifies LP-X methods, evaluation dimensions, and associated metrics. The model could be further extended to broader XAI methods. As a proof-of-concept, we instantiate the proposed EO extension with LP-DIXIT, a user-aware algorithmic explanation evaluation method, demonstrating the ontology's ability to address the targeted problem. Furthermore, we draw a solution for exploiting the semantic model, besides for the annotation and retrieval of different evaluation approaches based on multiple dimensions, but also for automating/operationalizing LP-Xs, given interest dimensions. The paper offers a view towards the foundation for a unified evaluation of post-hoc LP-Xs, and drafts the ground for automated user-centric assessment workflows.

Keywords

explainable AI, post-hoc explanations, link prediction, explanation evaluation, ontologies

1. Introduction

Knowledge Graphs (KGs) are formal machine-processable representations of knowledge consisting of entities (nodes) and binary relations (edges) [1]. Despite their proven utility, KGs are often noisy and/or incomplete, as they come as a result of a complex building process. Hence, Link Prediction (LP) methods, aiming at predicting missing facts, are leveraged for completing KGs. Mostly, LP tasks are solved by *Knowledge Graph Embeddings* (KGEs) models, which showed good performance [2]. Nevertheless, their black-box nature has raised the need for (*post-hoc*) explanations, especially in fields where LP may imply critical decisions, such as finance or pharmacology. For example, LP can be used to find new targets for existing drugs, reducing drug development costs. LP eXplanations (LP-X) would then clarify why and how predictions are made, enhancing trust and helping stakeholders to make informed decisions.

XAI-KRKG@ECAI25: First International ECAI Workshop on eXplainable AI, Knowledge Representation and Knowledge Graphs, October 25–30, 2025, Bologna, Italy

[†]The authors contributed equally.

lbalbi@lasige.di.fc.ul.pt (L. Balbi); felix.bindt@wur.nl (F. Bindt); katja.breitenfelder@ibp.fraunhofer.de (K. Breitenfelder); riccardo.campi@polimi.it (R. Campi); jitse.desmet@ugent.be (J. De Smet); claudia.damato@uniba.it (C. d'Amato)
http://lbalbi.github.io/ (L. Balbi); https://github.com/FPieter (F. Bindt); https://github.com/k-breitenfelder (K. Breitenfelder); https://jitsedesmet.be/ (J. De Smet)
0000-0002-1116-3476 (L. Balbi); 0000-0002-2087-8334 (F. Bindt); 0000-0001-7765-3438 (K. Breitenfelder); 0000-0003-0716-7631 (R. Campi); 0009-0002-6513-5013 (J. De Smet); 0000-0002-3385-987X (C. d'Amato)

LP-X methods for KGs [3, 4] often adopt evaluation protocols and metrics tailored to specific LP algorithms and benchmarks, with limited emphasis on enabling systematic evaluation and comparison among them. Explanations are sometimes assessed across various dimensions, such as their impact on predictive task performance, their usefulness to users, and their overall clarity. This hampers their reproducibility and makes it difficult to identify their trade-offs, determine best practices, and benchmark new ones against established baselines.

An appealing solution involves using ontologies to broadly model LP-X methods and their evaluations. By representing them into a shared conceptual space, such conceptualisation may not only categorise existing approaches but also drive to an operationalised design of a systematic and unified evaluation protocol with potential to be expanded to novel/additional evaluation dimensions. This would allow users to assess LP-X methods and evaluation protocols across domains, model types, and explanation styles, while covering all evaluation perspectives.

One promising resource is the Explanation Ontology (EO) [5], which describing XAI methods in terms of their inputs, outputs, and underlying assumptions, but lacks any formalisation of their evaluation processes and protocols. As such, without any conceptualisation for defining and selecting evaluation dimensions and corresponding metrics, it remains challenging to make a comprehensive and consistent comparison of these methods.

In this position paper, we argue that a unified, ontology-driven approach can overcome this fragmentation. We further draw a solution that sees the extension of the EO to define LP-Xs evaluation dimensions jointly with corresponding metrics, with the final goal of providing a comprehensive conceptualisation for automatising evaluating explanations, particularly coming from post-hoc LP-X solutions. Indeed, encoding explanation methods, metrics, and the evaluation protocol within the same semantic model (i.e., a structured knowledge representation like an ontology) facilitates retrieving and applying the most appropriate metrics for any given dimension for evaluating post-hoc explanation solutions, allowing comparability, reproducibility, and coverage across LP scenarios and beyond.

We frame two Research Questions (RQs) to support our position:

> **RQ 1.** What are the dimensions for assessing/evaluating post-hoc LP explanations?
>
> **RQ 2.** Can a semantic model be adopted to realize a unified automated framework for evaluating different LP-Xs methods?

RQ 1 lays the groundwork by identifying and organising the essential evaluation dimensions. **RQ 2** tests the hypothesis that an ontology-driven automatised system would be capable of employing the right evaluation methods and settings across diverse dimensions.

The paper is organised as follows: Sect. 2 provides an overview of post-hoc LP-X methods and evaluation protocols. Sect. 3 introduces some key notions for our proposal. Sect. 4 outlines our suggested direction and design of the solution Sect. 5 showcases a proof-of-concept to validate the solution. Sect. 6 recaps our position and suggestions and outlines directions for future research.

2. State of the Art

In this work, we specifically target post-hoc LP-X solutions. As such, in this section, we survey the main state-of-the-art in this direction and the class of metrics adopted for their evaluation.

Post-hoc LP-X methods differ in (1) the *form of explanation* they produce (e.g., facts, paths, rules, subgraphs), (2) the *compatibility with underlying KGE models*, and (3) the *evaluation protocol* applied - often without standardisation across studies. Early approaches [6] generate single-fact explanations via perturbations or influence functions, but are typically limited to specific model types. Successive solutions [7, 8] enhance flexibility by introducing post-training modules applicable to most embedding models. Other works extract relevant sets of facts or neighborhood subgraphs, e.g. Baltatzis and Costabello [9] employs knowledge distillation on sampled subgraphs, while Zhao et al. [10] identifies substructures based on information gain. Ma et al. [11] uses greedy search to isolate subgraphs most relevant to a prediction. Broader techniques extend beyond (set of) triples. Amador-Domínguez et al. [12] outputs ontological axioms or factual triples, supported by template-based natural language generation. Betz et al. [13] uses adversarial abduction over learned rules while Ismaeil et al. [14] extracts interpretable features from embedding vectors for downstream tasks. Path-based approaches [6, 15] identify semantically similar paths using relation and entity similarity. Other solutions [16] adopt rule mining evaluated through classification performance.

Existing protocols for evaluating LP-X typically fall under one of the following aspects [17, 18]:

1. **Functionally grounded**: Assess LP-X without human subjects, by probing over the KG and the model's scoring. Their measures capture model response, either on changes in model decision (faithfulness), agreement with a surrogate model of the explanation (fidelity), similarity of explanations across perturbations (stability) and axiom violations or entailments (consistency), among others.
2. **Human grounded**: Evaluation over tasks that probe comprehensibility and practical usefulness, with focus on measures of accuracy, time-to-decision, preference proportions, trust rate and user-agreement among others.
3. **Application grounded**: With human-involvement for explanation evaluation in the context of a given domain application, with measures that capture whether explanations actually improve decisions and workflows (e.g., utility, expert acceptance rate).

Each measure has a metric score it can be tied to to operationalize its evaluation aspect (e.g., measure of faithfulness - necessity/sufficiency rate).

Despite this rich landscape, evaluation approaches have typically been applied only to a subset of these metrics in isolation, without a standardized and unified protocol for selecting and combining them. Similar to how human-grounded and application-grounded metrics have been split, recent work highlights the benefits of using large language models as a proxy for human-grounded metrics [19, 20].

3. Basics

This section surveys the various dimensions of the state of the art for evaluating explanations [21] and their relevance to the general evaluation of LP-X. Evaluation dimensions that are not deemed relevant in the context of this work are marked†. Evaluation dimensions that require some, to be further defined, underlying distance metric are marked differently$^{\triangle}$. We conclude this section by arguing on the need of actual computable metrics that are currently mostly missing.

Functionally grounded

Faithfulness measures the accuracy of the explanation regarding the prediction. With respect to LP-X this means that if the explanation of triple ‘x‘ is a set of triples ‘Y‘, when you remove a subset of ‘Y‘, the LP-model no longer predicts triple ‘x‘. As such, the LP-X model is faithful to the LP model, since a removal of the justification does indeed mean that the prediction would not have been made. **Location accuracy**†measures the ability for an explanation model to localize the explanation correctly with respect to some points of interests within the ground truth. Since the concrete meaning of points of interest is ill-defined within the context of LP-X, we ignore this dimension in the remainder of our work. **Completeness** measures how much of the actual reason is covered by the explanation, necessitating a ground truth. **Overlap**†is a dimension specifically targeting rule-based systems. Since we want to operate in the general case, we will not consider it further. **Accuracy**†a metric usable when you use a surrogate model to provide explainability. Again, since we want to model the general evaluation, we will not consider this metric further. **Architectural complexity**†and **algorithmic complexity**†are two metrics that are hard, if not impossible, to compare across different explainability settings. **Stability**$^{\triangle}$measures the stability of an explanation given a change in the underlying data. Within LP-X, this would mean that you measure how different your explanation is given an independent modification in the underlying graph the explanator can use. **Consistency**$^{\dagger\triangle}$measures the change of the explanator given a small change in the to be explained information. Since the general LP problem does not predict literals e.g. strings, dates, numbers, etc), this is not relevant in the general setting because a ’slightly different’ prediction does not exist. Each difference in relation is similarly different because it talks about a different resource. **Sensitivity**$^{\triangle}$in the context of LP-X measures how different the explanation of the X-model is when provided with a different input triple to explain. Basically, it punishes models that would always provide the same explanation regardless of what they need to explain. **Expressiveness** measures the level of detail used by the X-model within the formal model (e.g. triple count for LP-X).

Human grounded

Interoperability/ complexity: measures how well a user can make a mental model of the explanation. **Effectiveness** is the accuracy of human thinking for the predicted triple after seeing the explanation. It functions as a proxy for *interoperability*. **Time efficiency**†measures how long it takes for a user to build a viable mental model. Since there is no explanation feedback loop, *information amount* acts as a proxy to this measurement. **Degree of understanding**†: measures in interactive contexts the current status of understanding. This is not generally applicable in post-hoc LP-X, since the model will typically explain once, and not be asked to generate more detailed explanations. **Information Amount** is the amount of information conveyed through the explanation. It could be measured by something like 'triple count of the explanation' but this is incomplete since the explanation triple could, for example, have a singleton property, in which case the human in the loop is still getting a lot of information in reality.

Application grounded

Satisfaction measures how content the explainee is with the system. A well-known metric is for example the System Usability Scale (SUS) score [22]. **Persuasiveness** measures how persuasive the generated explanations are. Whether high persuasiveness is a good or bad thing mostly depends on the context - the extremes are often avoided. **Improvement of human judgment** assesses to what degree the user gets to trust the system. Correct explanations should be trusted more. **Improvement of human-AI system performance**, measures the total system, why you want the link predicted, who wants it and whether the explanation they get improves the situation. **Automation capability** is a metric that tries to uncover whether a human actually spends less time identifying missing relations/ links. It asks to what extend the overall system reduces manual labour. **Novelty** measures whether the links predicted, and the provided explanations, highlight novel discoveries. An example of a novel discovery would be the prediction and justification of a triple ':somePill a :cureToCancer'.

We conclude by highlighting that even though functionally grounded evaluation dimensions focus on enabling automated testing, only a limited set of dimensions are actually applicable in the context of LP-X. Specifically, only *faithfulness*, *completeness* and *expressiveness*. Moreover, both completeness and expressiveness present challenges: completeness relies on well-defined ground truth, while expressiveness lacks robust evaluation metrics. This analysis underscores the need for better functionally grounded dimensions and metrics, while also motivating the use of LLMs as a proxy for both human and application-grounded dimensions, as described by Barile et al. [19].

4. Semantic Framework

The proposed semantic evaluation framework is grounded on a conceptualisation that describes LP-X evaluation methods and settings. This framework will be used for enabling systems/agents to select and execute evaluation protocols in a unified manner and potentiate the automation and standardisation of LP-X evaluations. Specifically, given the EO [5], we showcase its extension to serve as the exoskeleton of an evaluation system (with evaluation terms that are missing in EO), by adding high-level concepts of LP-X evaluation into ontology classes and properties.

In Sect. 4.1 we describe the EO and in Sect. 4.2 we illustrate its extension to support our proposal of a semantic-driven solution to build a unified and automated LP-X evaluation framework. Sect. 4.3 drafts the envisioned solution for automatizing LP-X evaluation.

4.1. Explanation Ontology

The EO [5] is a general purpose semantic model to represent and connect user-centric explanations to the underlying data and knowledge with the end-goal of making model recommendations more explainable. As illustrated in Fig. 2 reported in Annex A.1, the ontology is organised into three conceptual layers: the *User Layer*, that models user-centric goals and preferences (e.g. ExplanationGoal and UserProfile); the *Interface Layer* that captures explanation modalities and presentation formats (e.g. ExplanationModality); and the *System Layer* that describes the internal representation of explainers, linking to data sources, models and provenance (e.g. ExplanationMethod, SystemRecommendation).

Although its current format covers explanation generation thoroughly and supports a broad range of state-of-the-art explainer methods, EO still lacks formal constructs to describe the evaluation of (post-hoc) LP-X solutions. For this reason, we propose an extension of the EO in Sect. 4.2.

4.2. Extending the Explanation Ontology

One common strategy to guide ontology development and enrichment, and to assess the quality of an ontology with respect to a specific application is using competency questions (CQs), i.e. questions formulated in natural language, representing the requirements to be answered using data structured according to the ontology [23]. Since these are questions with established or verifiable answers, they also function as a form of content validation techniques to determine if the ontology fits the requirements and is structurally sound.

We therefore propose the following CQs as a first step to assess the capability of our targeted extended EO (EEO) to support the proposed evaluation system:

> **CQ1**: Which measure(s) are available for an evaluation aspect/dimension or set of aspects YY?
>
> **CQ2**: Which method(s) should be used to perform an explanation evaluation on the dimension/ set of aspects YY and set of measures ZZ?

The extension of EO[1] (EEO) follows a top-down approach by defining the high-level concepts of LP-X evaluation into classes and properties and then specifying into the more specific concepts they may contain. These concepts were selected from the dimensions illustrated in Sect. 3.

We list a minimal set of currently missing high-level classes and properties that we deem as essential for encoding the evaluation process in the ontology and for allowing the instantiation of LP-X methods and procedures. The full list of classes and properties is reported and documented in Appendix A.3, while Figure A.2 of Appendix A.2 illustrates the schema that results from extending the EO. At a high level, we introduce classes for Explanation Evaluation, Evaluation Measure and Quantitative Measure that in turn contain subclasses pertaining to the dimensions, measures and metrics identified in the SOTA [21] (and summarized in Sect. 3). We deem these classes as necessary but possibly extensible for modeling further LP-X/XAI evaluation dimensions. We have therefore adopted Protegé 5.6.3 to introduce the listed classes in the EO within a novel `eeo:` namespace corresponding to the EEO. These classes as well as the newly created object properties resulted logically consistent with the existing EO. The EEO retains full backward compatibility with EO while providing the semantic hooks needed to model evaluation workflows. Sect. 5 illustrates the instantiation of the EEO, showcasing the support of existing (and newly developed) LP-X solutions.

4.3. Automated LP-X explanation evaluation

The proposed EEO should enable the automation of evaluation protocols for LP-X models. To support this, the EEO could be integrated in a holistic (agenting) solution that, given a user-inputted LP problem, and LP-X method(s) output data, allows: 1) the querying of EEO for different explanation evaluation dimension(s); 2) the collection of LP-X evaluation methods and metrics supporting the queried dimension(s) and input LP problem; and 3) the automated execution of the LP-X evaluation protocols that support those methods (see Figure 1). In particular, the agent should be able to recognise the LP-X method and explanation type provided by the user and translate them into a SPARQL query to retrieve all relevant dimensions, metrics and protocols from the EEO. The query results should serve as input to the evaluation module along with the original explanation data. The agent would then return a complete LP-X evaluation report to the user, with evaluation methods' results organized along the

[1] The extension of EO is avaiable at https://doi.org/10.5281/zenodo.15658539

different dimensions. A unified, end-to-end system of this kind could not only reduce the cognitive burden for researchers but also standardize LP-X evaluation protocols.

5. Proof-of-Concept

Our proof-of-concept covers validating EEO and verifying its compliance with predefined requirements through answering the CQs established in Sect. 4 formalising corresponding SPARQL queries. For this purpose, we instantiate our ontology considering LP-DIXIT [19], an algorithmic and user-centric LP-X evaluation solution. The method provides an evaluation dimension and measure relevant for testing this approach: it assesses the user-grounded aspect, over the utility of explanations, with a quantitative *forward simulatability variation* metric on the improvement of a user's prediction accuracy when given an explanation. The method was used to specifically populate the EEO through the `eeo:EvaluationMethod` class and annotated to the pertaining Evaluation Measure class, a subclass of `eeo:EvaluationMeasure`.

To answer our previously defined CQs, we designed SPARQL queries (listing 1 and 2) specified for the aspect and measure evaluated in the LP-DIXIT method. Through these we demonstrate that the utility of explanations - reflected by user agreement - can be captured within the user perspective dimension of our semantic model. Furthermore, not only do they provide the classes linking dimensions and metrics for evaluating explanations (listing 1), but also return LP-DIXIT as an instance for describing the methodology (listing 2), confirming successful mapping.

6. Position Summary

Despite the emerging works on evaluating current XAI methods on their different characteristics and along several axes of explainability, there is still a lack of systematized benchmarks and of unified comparison systems between the different evaluation approaches.

After a broad review of the SOTA in XAI evaluation, we identified several of the evaluation dimensions/aspects that have been approached so far and propose an ontology-driven evaluation system for post-hoc explanations in KG LP tasks. Our proposed system maps current methodologies to their evaluation protocols, aspects and measures to enable structured and unified evaluation workflows.

To do so, we suggest extending an existing ontology, EO, to incorporate XAI evaluation constructs. Furthermore, we provide a proof-of-concept approach to validate the EEO and its incorporation into the proposed evaluation framework via answering an initial pair of CQs. These CQs are able to obtain protocols and metrics for evaluating post-hoc LP-X methods in the dimension of user-experience. Moreover, we propose an automated system integrating the EEO together with an agent to relieve workload and standardize evaluations.

To conclude, we encourage researchers to adopt XAI ontology-driven or similar approaches to building universal and systematic evaluation frameworks that allow for consistent comparison and benchmarking across XAI methods.

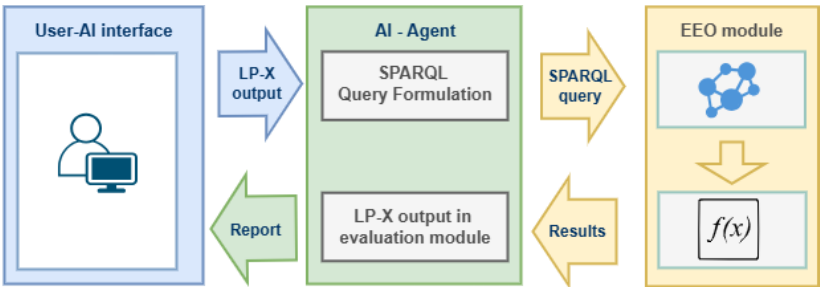

Figure 1: Data pipeline for automated LP-X evaluations using the EEO. It consists of: (1) the user providing LP-X outputs, (2) an AI-agent which is able to understand, translate, and implement the LP-X and EEO results, and (3) the EEO module where explanations and metrics are stored.

Acknowledgments

Laura Balbi is funded by FCT through grant number 2024.01208.BD and partially funded by the EU through the KATY project (grant agreement No 101017453). Jitse De Smet is a predoctoral fellow of the Research Foundation – Flanders (FWO) (1SB8525N). Felix Bindt is funded by the Strategic Program of the Netherlands National Institute for Public Health and the Environment (RIVM) (No S/133030/01). Katja Breitenfelder acknowledges the support of Fraunhofer IBP, Valley, Germany, which contributed to the completion of this research. Riccardo Campi is funded by the European Commission's Horizon Europe project ENERGENIUS, Project ID 101160720. Claudia d'Amato was partially supported by project *FAIR - Future AI Research* (PE00000013), spoke 6 - Symbiotic AI (https://future-ai-research.it/) under the PNRR MUR program funded by the European Union - NextGenerationEU, and by PRIN project *HypeKG - Hybrid Prediction and Explanation with Knowledge Graphs* (Prot. 2022Y34XNM, CUP H53D23003700006) under the PNRR MUR program funded by the European Union - NextGenerationEU.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (2022) 71:1–71:37. URL: https://doi.org/10.1145/3447772. doi:10.1145/3447772.

[2] J. Cao, J. Fang, Z. Meng, S. Liang, Knowledge graph embedding: A survey from the perspective of representation spaces, ACM Comput. Surv. 56 (2024) 159:1–159:42. URL: https://doi.org/10.1145/3643806. doi:10.1145/3643806.

[3] P. Pezeshkpour, Y. Tian, S. Singh, Investigating robustness and interpretability of link prediction via adversarial modifications, in: 1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019, 2019. URL: https://openreview.net/forum?id=Hkg7rbcp67.

[4] H. Zhang, T. Zheng, J. Gao, C. Miao, L. Su, Y. Li, K. Ren, Data poisoning attack against knowledge graph embedding, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 4853–4859. URL: https://doi.org/10.24963/ijcai.2019/674. doi:10.24963/IJCAI.2019/674.

[5] S. Chari, O. Seneviratne, M. F. Ghalwash, S. Shirai, D. M. Gruen, P. Meyer, P. Chakraborty, D. L. McGuinness, Explanation ontology: A general-purpose, semantic representation for supporting user-centered explanations, Semantic Web 15 (2024) 959–989. URL: https://doi.org/10.3233/SW-233282. doi:10.3233/SW-233282.

[6] W. Zhang, B. Paudel, W. Zhang, A. Bernstein, H. Chen, Interaction embeddings for prediction and explanation in knowledge graphs, in: J. S. Culpepper, A. Moffat, P. N. Bennett, K. Lerman (Eds.), Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019, ACM, 2019, pp. 96–104. URL: https://doi.org/10.1145/3289600.3291014. doi:10.1145/3289600.3291014.

[7] A. Rossi, D. Firmani, P. Merialdo, T. Teofili, Explaining link prediction systems based on knowledge graph embeddings, in: Z. G. Ives, A. Bonifati, A. E. Abbadi (Eds.), SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022, ACM, 2022, pp. 2062–2075. URL: https://doi.org/10.1145/3514221.3517887. doi:10.1145/3514221.3517887.

[8] R. Barile, C. d'Amato, N. Fanizzi, Explanation of link predictions on knowledge graphs via levelwise filtering and graph summarization, in: A. Meroño-Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, P. Lisena (Eds.), The Semantic Web - 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part I, volume

14664 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 180–198. URL: https://doi.org/10.1007/978-3-031-60626-7_10. doi:10.1007/978-3-031-60626-7_10.

[9] V. Baltatzis, L. Costabello, Kgex: Explaining knowledge graph embeddings via subgraph sampling and knowledge distillation, in: S. Villar, B. Chamberlain (Eds.), Learning on Graphs Conference, 27-30 November 2023, Virtual Event, volume 231 of *Proceedings of Machine Learning Research*, PMLR, 2023, p. 27. URL: https://proceedings.mlr.press/v231/baltatzis24a.html.

[10] D. Zhao, G. Wan, Y. Zhan, Z. Wang, L. Ding, Z. Zheng, B. Du, KE-X: towards subgraph explanations of knowledge graph embedding based on knowledge information gain, Knowl. Based Syst. 278 (2023) 110772. URL: https://doi.org/10.1016/j.knosys.2023.110772. doi:10.1016/J.KNOSYS.2023.110772.

[11] T. Ma, X. Song, W. Tao, M. Li, J. Zhang, X. Pan, J. Lin, B. Song, X. Zeng, Kgexplainer: Towards exploring connected subgraph explanations for knowledge graph completion, CoRR abs/2404.03893 (2024). URL: https://doi.org/10.48550/arXiv.2404.03893. doi:10.48550/ARXIV.2404.03893. arXiv:2404.03893.

[12] E. Amador-Domínguez, E. Serrano, D. Manrique, Geni: A framework for the generation of explanations and insights of knowledge graph embedding predictions, Neurocomputing 521 (2023) 199–212. URL: https://doi.org/10.1016/j.neucom.2022.12.010. doi:10.1016/J.NEUCOM.2022.12.010.

[13] P. Betz, C. Meilicke, H. Stuckenschmidt, Adversarial explanations for knowledge graph embeddings, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, ijcai.org, 2022, pp. 2820–2826. URL: https://doi.org/10.24963/ijcai.2022/391. doi:10.24963/IJCAI.2022/391.

[14] Y. Ismaeil, D. Stepanova, T. Tran, H. Blockeel, Feabi: A feature selection-based framework for interpreting KG embeddings, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I, volume 14265 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 599–617. URL: https://doi.org/10.1007/978-3-031-47240-4_32. doi:10.1007/978-3-031-47240-4_32.

[15] C. d'Amato, P. Masella, N. Fanizzi, An approach based on semantic similarity to explaining link predictions on knowledge graphs, in: J. He, R. Unland, E. S. Jr., X. Tao, H. Purohit, W. van den Heuvel, J. Yearwood, J. Cao (Eds.), WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence, Melbourne VIC Australia, December 14 - 17, 2021, ACM, 2021, pp. 170–177. URL: https://doi.org/10.1145/3486622.3493956. doi:10.1145/3486622.3493956.

[16] N. A. Krishnan, C. R. Rivero, A model-agnostic method to interpret link prediction evaluation of knowledge graph embeddings, in: I. Frommholz, F. Hopfgartner, M. Lee, M. Oakes, M. Lalmas, M. Zhang, R. L. T. Santos (Eds.), Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023, ACM, 2023, pp. 1107–1116. URL: https://doi.org/10.1145/3583780.3614763. doi:10.1145/3583780.3614763.

[17] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[18] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115. URL: https://doi.org/10.1016/j.inffus.2019.12.012. doi:10.1016/J.INFFUS.2019.12.012.

[19] R. Barile, C. d'Amato, N. Fanizzi, LP-DIXIT: evaluating explanations for link predictions on knowledge graphs using large language models, in: G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, Z. H. Huang, E. Yom-Tov (Eds.), Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025, ACM, 2025, pp. 4034–4042. URL: https://doi.org/10.1145/3696410.3714667. doi:10.1145/3696410.3714667.

[20] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, in: A. Oh, T. Nau-

mann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

[21] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Data Min. Knowl. Discov. 38 (2024) 3043–3101. URL: https://doi.org/10.1007/s10618-022-00867-8. doi:10.1007/S10618-022-00867-8.

[22] J. Brooke, SUS – a quick and dirty usability scale, Taylor & Francis, 1996, pp. 189–194.

[23] C. Bezerra, F. Freitas, F. Santana, Evaluating ontologies with competency questions, in: 2013 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Atlanta, Georgia, USA, 17-20 November 2013, Workshop Proceedings, IEEE Computer Society, 2013, pp. 284–285. URL: https://doi.org/10.1109/WI-IAT.2013.199. doi:10.1109/WI-IAT.2013.199.

Appendix

A.1. High Level Picture of the Explanation Ontology

This section reports figure 2 providing the high-level description of the EO [5].

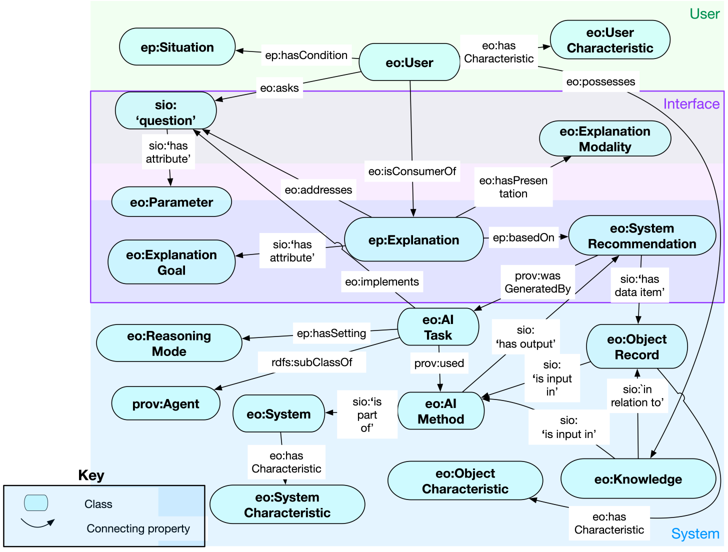

Figure 2: A conceptual overview of EO. Color shading in this diagram is used to depict the separation between user, system, and interface attributes.

A.2. High Level Picture of the Extended Explanation Ontology

This section reports the figure 3 providing the high-level description of the EEO ontology.

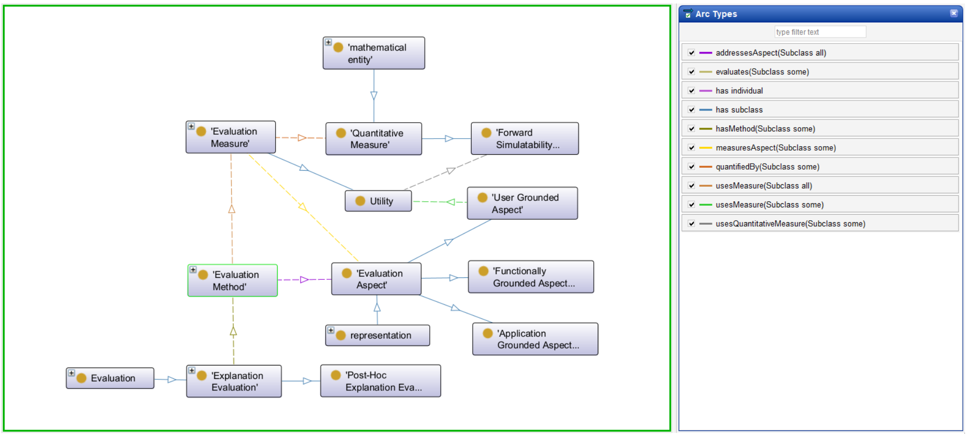
'mathematical entity'
'Evaluation Measure'
'Quantitative Measure'
'Forward Simulatability...
Utility
'User Grounded Aspect'
'Evaluation Method'
'Evaluation Aspect'
'Functionally Grounded Aspect...
representation
'Application Grounded Aspect...
Evaluation
'Explanation Evaluation'
'Post-Hoc Explanation Eva...
Arc Types
type filter text
☑ addressesAspect(Subclass all)
☑ evaluates(Subclass some)
☑ has individual
☑ has subclass
☑ hasMethod(Subclass some)
☑ measuresAspect(Subclass some)
☑ quantifiedBy(Subclass some)
☑ usesMeasure(Subclass all)
☑ usesMeasure(Subclass some)
☑ usesQuantitativeMeasure(Subclass some)

Figure 3: High-level description of the EEO to include the LP-X Evaluation component.

A.3. Class Definition

Listed High-level Classes

`eo:Evaluation` - Denoting an assessment activity.

`eo:ExplanationEvaluation` - Represents an evaluation with an explanation as input.

`eo:PostHocExplanationEvaluation` - Sub-class of Explanation Evaluation that denotes evaluation of XAI approaches employed over ML after training.

`eo:EvaluationMethod` - class describing the evaluation procedure with subclasses that capture local/global settings and task types.

`eo:EvaluationAspect` - top-level class for the dimension or quality being assessed (Function-grounded, Application-grounded or User-grounded).

`eo:FunctionallyGroundedAspect` - Aspects measurable without human interference, e.g. fidelity, monotonicity.

`eo:ApplicationGroundedAspect` - Aspects stemming from evaluation over domain-expert tasks.

`eo:UserGroundedAspect` - Aspects that require a human study or a simulation study with an agent component, e.g. on interpretability.

`eo:EvaluationMeasure` - A measure or metric used to assess explanations on a given evaluation aspect.

`eo:QuantitativeMeasure` - defining a concrete evaluation metric of quantitative nature (e.g. accuracy, recall, information content score).

`eo:EvaluationResult` - top-level class that represents the outcome of an Evaluation Method, linking evaluation measures to their quantitative values.

`eo:EvaluationAgent` - defining the actor conducting the evaluation, of human or automated nature.

A.4. Property Definition

List of added Object Properties between high-level classes

`eeo:evaluatesExplanation` *Domain*: `eeo:ExplanationEvaluation` *Range*: `eeo:Explanation`

`eeo:hasMethod` *Domain*: `eeo:ExplanationEvaluation` *Range*: `eeo:EvaluationMethod` *Axiom*: `eeo:ExplanationEvaluation SubClassOf (eeo:hasMethod min 1 eeo:EvaluationMethod)`

`eeo:addressesAspect` *Domain*: `eeo:EvaluationMethod` *Range*: `eeo:EvaluationAspect` *Axiom*: `eeo:EvaluationMethod SubClassOf (eeo:addressesAspect min 1 eeo:EvaluationAspect)`

`eeo:usesMeasure` *Domain*: `eeo:EvaluationMethod` *Range*: `eeo:EvaluationMeasure` *Axiom*: `eeo:EvaluationMethod SubClassOf (eeo:usesMeasure min 1 eeo:EvaluationMeasure)`

`eeo:hasAgent` *Domain*: `eeo:EvaluationMethod` *Range*: `eeo:EvaluationAgent` *Axiom*: `eeo:EvaluationMethod SubClassOf (eeo:hasAgent min 1 eeo:EvaluationAgent)`

`eeo:producesResult` *Domain*: `eeo:EvaluationMethod` *Range*: `eeo:EvaluationResult` *Axiom*: `eeo:EvaluationMethod SubClassOf (eeo:producesResult min 1 eeo:EvaluationResult)`

`eeo:measuresAspect` *Domain*: `eeo:EvaluationMeasure` *Range*: `eeo:EvaluationAspect`

`eeo:quantifiedBy` *Domain*: `eeo:EvaluationMeasure` *Range*: `eeo:QuantitativeMeasure`

List of added Data Properties

`eeo:hasValue` *Domain*: `eeo:QuantitativeMeasure` *Range*: `xsd:decimal`

`eeo:hasUnit` *Domain*: `eeo:QuantitativeMeasure` *Range*: `xsd:string`

A.5. Proof-of-Concept SPARQL queries

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX eeo: <https://purl.org/heals/eo#>

SELECT DISTINCT ?EvalMeasure ?Metric
WHERE {
    VALUES (?EvalAspect) {(Aspect_IRI)}

    ?Evalmethod rdf:type eeo:EvaluationMethod .
    ?Evalmethod eeo:hasAspect ?EvalAspect .
    ?Evalmethod eeo:usesMeasure ?EvalMeasure .
    ?EvalMeasure eeo:measuresAspect ?EvalAspect .
    ?EvalMeasure eeo:usesQuantitativeMeasure ?Metric .
}
```

Listing 1: SPARQL query for proof-of-concept answering CQ1, "Which measure(s) are available for a specified evaluation aspect or set of aspects YY?"

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX eeo: <https://purl.org/heals/eo#>

SELECT DISTINCT ?Evalmethod ?Metric
WHERE {
    VALUES (?EvalMeasure) {(Utility_IRI)}
    ?Evalmethod rdf:type eeo:EvaluationMethod .
    ?Evalmethod eeo:hasAspect ?EvalAspect .
    ?Evalmethod eeo:usesMeasure ?EvalMeasure .
    ?EvalMeasure eeo:measuresAspect ?EvalAspect .
    ?EvalMeasure eeo:usesQuantitativeMeasure ?Metric .
}
```

Listing 2: SPARQL query for proof-of-concept answering CQ2, "Which method(s) should be used to perform an explanation evaluation on the dimension/set of aspects YY and set of measures ZZ?"