Taxonomy of Anomaly Types in Knowledge Graphs

Asara Senaratne[1(✉)], Peter Christen[2], Pouya Omran[2], and Graham Williams[2]

[1] College of Science and Engineering, Flinders University, Tonsley, Australia
asara.senaratne@flinders.edu.au
[2] School of Computing, Australian National University, Canberra, Australia
{peter.christen,p.g.omran,graham.williams}@anu.edu.au

Abstract. Detecting anomalies in Knowledge Graphs (KG) is a challenging task as the patterns of anomalies are unpredictable, unknown, diverse, likely rare, and often with no ground truth labels available. Hence, it is important to identify the types of such anomalies occurring in a KG, so domain experts can adopt measures to prevent anomalies occurring during KG construction, or remove anomalies from already constructed KGs, while also discovering knowledge. In such a process we can obtain a classification among these identified anomalies such that we know what anomalies are to be forwarded to domain experts for correction, and what can be corrected via automatic or semi-automatic techniques. However, to the best of our knowledge, there is no such predefined classification of possible common anomalies that could arise in a KG, which we could directly use to support anomaly classification. Hence, in this paper, we propose a taxonomy of possible anomaly types that can occur in a KG using the real-world KGs YAGO-1, DSKG, Wikidata and KBpedia.

Keywords: Anomaly classification · Knowledge Graph refinement · Data quality · Anomaly grouping

1 Introduction

Data quality management methods available for traditional relational data cannot be applied on Knowledge Graphs (KG) due to either of the following reasons [13]. First, unlike relational data, KGs are semi-structured and often do not come with a schema to specify the integrity and semantics of data. This heterogeneity of data and the flexibility of schemas make structures more complex. Second, the semantic web and KGs are usually built on the Open World Assumption [3], where a statement not included in the KG is assumed to be wrong or absent. So, it is not easy to distinguish a wrong fact from a missing fact. Lastly, due to the scale of real-world KGs which is typically beyond the capacity of existing data quality methods to process, a direct application of such techniques can be time consuming. Therefore, new solutions for KG quality management should emerge gradually [13]. The introduction of taxonomies

E. Curry et al. (Eds.): ESWC 2025, LNCS 15832, pp. 151–155, 2026.
https://doi.org/10.1007/978-3-031-99554-5_28

Table 1. Anomaly types involving entity-based literals with examples extracted from YAGO-1, DSKG, Wikidata, and KBpedia.

Triples involved	Anomaly type	Anomaly name	Example	Anomaly source	Possible correction
Single triple	Missingness	Missing subject	<, bornIn, London>	Storage file	Link prediction
		Missing predicate	<9thWonder, TheDreamMerchantVol2>	Storage file	Link prediction
		Missing object	<Distribution/100150, byteSize, >	Storage file	Link prediction
	Incorrectness	Incorrect predicate	<MarlaMaples, hasChild, DonaldTrump>	Graph	Link prediction
	Inconsistency	Invalid predicate	<Stavisky, isLocatedIn, France>	Graph	Link prediction
Multiple triples	Ambiguity	Entity ambiguity	<JoshGracin, originatesFrom, JoshGracin>	Graph	Entity disambiguation
			<Marcelona, bornIn, Mozambique>	Graph	Entity disambiguation
			<Marcelona, hasSuccessor, Mozambique>		
		Predicate ambiguity	<Ain'tTooProudToBeg, isOfGenre, rock>	Graph	Human evaluation, or KG re-engineering
			<Ain'tTooProudToBeg, isOfGenre, music>		
			<Ain'tTooProudToBeg, isOfGenre, popularMusic>		
	Contradictions	Contradicting facts	<DonaldTrump, marriedTo, MarlaMaples>	Graph	Remove incorrect RDF entry
			<MarlaMaples, hasChild, DonaldTrump>		
	Unusual	Rare entity	Entity *9thWonder* has only one fact	Graph	No correction required.
		Prolific entity	Entity *Dataset/410* has 11 *createdBy* links	Graph	No correction required.
	Redundancies	Redundant facts	<AMGRapper, produced, BettaHaveMoney>	Graph	Human evaluation
			<AMGRapper, produced, BettaHaveMoney2001>		
	Duplicates	Duplicate facts	<LizBeth, bornIn, Mexico>	Storage file	Remove duplicates
			<LizBeth, bornIn, Mexico>		

is one such effort made by the research community. A scheme of classification, usually a hierarchical classification, in which things are organized into groups, types, or classes is named a taxonomy [2].

Contribution: Due to the absence of a taxonomy of anomaly types for KGs, we propose TAXO (TAXOnomy of anomaly types in KGs), a unification of an extensive set of anomaly types in KGs that we can discover either by analyzing KG data storage files such as Notation3 (N3)[1], or by representing data as a graph. TAXO can support domain experts to prevent the identified anomalies from occurring in new KGs, and to broaden their view [10] when developing algorithms to identify anomalies in existing KGs. The main contribution of our work is towards the enhancement of data quality thereby generating enriched KGs. A unifying view of anomalies provides a solid foundation to understand the severity of these anomalies, discovers knowledge, and supports future research in this area.

2 Proposed Taxonomy

Based on our previous work in graph anomaly detection [7,9,11], we constructed TAXO as an effort to group and classify identified anomalies so that this taxonomy can act as a resource for KG enrichment tasks. TAXO considers possible anomaly types that can occur either in an RDF storage file such as Terse RDF Triple Language (Turtle)[2], and anomaly types that we can discover upon graph population. We primarily classify anomaly types based on the type of a triple, where we separate the anomalies occurring in triples of the form *entity-to-entity*, from the anomalies occurring in triples of the form *entity-to-literal*. This separation is useful as the anomaly detected and the method of correction differs based

[1] https://www.w3.org/TeamSubmission/n3/.
[2] https://www.w3.org/TR/turtle/.

Table 2. Anomaly types involving literal-based triples with examples extracted from YAGO-1, DSKG, Wikidata, and KBpedia.

Triples involved	Anomaly type	Anomalyname	Example	Anomaly source	Possible correction
Single triple	Missingness	Missing subject	<, hasDefinition, "A query language">	Storage file	Remove RDF entry
		Missing predicate	<Echidna, "Egg laying mammal">	Storage file	Link prediction
		Missing literal	<SQL, hasDefinition, "">	Graph	Link prediction
	Incorrectness	Incorrect literal	<Aristotle, bornOn, "380">	Graph	Human evaluation
			<DJShadow, created, "album">		
		Partially correct literal	<AliHewson, bornOn, "1961-##-##">	Graph	Human evaluation
	Inconsistency	Invalid predicate	<AMGAlbum, bornOn, "2001-11-25">	Graph	Human evaluation
Multiple triples	Redundancies	Redundant literals	<SQL, isa, "Programming Language">	Graph	Human evaluation
			<SQL, isa, "Programming-Language">		
			<Zucchini, altLabel, "fruit of zucchini plant">		
			<Zucchini, altLabel, "fruit of zucchini plants">		
			<Zucchini, altLabel, "fruit of zucchini">		
	Duplicates	Duplicate facts	<LizBeth, bornOn, "1948-04-20">	Graph	Remove duplicates
			<LizBeth, bornOn, "1948-04-20">		

on the triple type, and not every triple type has the same type of anomalies. For each triple type, we further classify anomalies based on the number of triples involved in the anomaly as single or multiple, as provided in Tables 1 to 3. A detailed discussion of this proposed taxonomy is available here [8].

2.1 Entity-Based Triples

In Table 1, we identify eight different types of anomalies that can occur involving a triple of the form *entity-to-entity*, where both its subject and object are entities. Anomalies such as a missing element, incorrectness, and type inconsistencies, usually involve a single triple. Whereas, anomalies such as ambiguity, contradictions, redundancies, and duplicates involve multiple triples. Interestingness can involve triples or entities that are classified as abnormal, but non-erroneous.

2.2 Literal-Based Triples

In Table 2, we identify six types of anomalies that can occur involving one or multiple triples of the form *entity-to-literal*, where the subject of such a triple is an entity, and the object is a literal. A literal can be a string, date, number, or hyperlink. Anomalies such as missing elements, or an incorrect, partially correct, and invalid triple usually involve a single triple, while redundancies and duplicates involve multiple triples. Similar to anomalies that involve either entity-based (such as the examples in Table 1) or literal-based triple (such as the examples in Table 2), there are anomalies that occur due to contradictions between triples from both these types. We name this category as mixed triples-based anomalies. As shown in Table 3, while multiple triples together form this anomaly, it is possible that all or one of the involved triples convey incorrect information, thus creating a contradiction.

Table 3. Anomaly types involving both entity-based and literal-based triples with an example extracted from YAGO-1.

Triples involved	Anomaly type	Anomaly name	Example	Anomaly source	Possible correction
Multiple triples	Contradiction	Mixed triples	<DonaldTrump, hasChild, DonaldTrumpJr.> <DonaldTrump, altLabel, "Donald Trump Jr.,">	Graph	Human evaluation

3 Literature Review

Researchers introduce taxonomies to present information about a particular knowledge area, or to support automation tasks by providing a unified understanding and categorization of terms used in a particular domain [6]. From computer science to psychology, and education, taxonomies have been used in every scientific domain to establish a common understanding among domain experts.

For example, Sutcliffe et al. [12] propose a taxonomy of error types for failure analysis and risk assessment. As failure in human computer systems can be due to many different causes, understanding these reasons behind design failures are particularly important for safety-critical systems. Lee et al. [5] propose a list of tasks required for graph visualization that has enough detail and specificity to be useful to designers who want to improve their systems, and to evaluators who want to compare graph visualization systems. Similarly, Zaveri et al. [14] present the results of a systematic review of approaches for assessing the quality of linked data. In particular, the authors unify and formalize commonly used terminologies across papers related to data quality and provide a comprehensive list of quality dimensions and metrics.

Therefore, taxonomies are an important resource to better understand issues, challenges, and trends in various domains [2]. For example, in the semantic web, work proposed by Breit et al. [1] provides a classification for machine learning-based semantic web systems which can be used as a template to analyse existing semantic web systems and to describe new ones. This template provides a controlled vocabulary for different building blocks of those systems. Furthermore, Gomez et al. [4] present a brief summary of previous work done on evaluating ontologies and the criteria used to evaluate and to assess ontologies. The authors also address the possible types of errors made when domain knowledge is structured in taxonomies in an ontology and in knowledge bases.

Although the importance of taxonomies has been identified decades ago [12], to the best of our knowledge, a taxonomy of anomaly types in KGs has not been proposed so far. We aim to address this research gap by introducing TAXO.

4 Conclusion

In this paper, we introduced TAXO, a taxonomy that classifies different types of anomalies in Knowledge Graphs (KGs). It groups anomalies by the type of triple involved and how many triples are needed to identify the anomaly. To make things clearer, TAXO includes examples from four real-world KGs and suggests

ways to fix each anomaly. TAXO can support KG enrichment and validation, and it can also be used to train Large Language Models for detecting anomalies in KGs. In the future, we plan to enhance TAXO by adding complexity levels for each anomaly type to better assist domain experts.

References

1. Breit, A., et al.: Combining machine learning and semantic web: a systematic mapping study. Comput. Surv. **313**, 1–41 (2023)
2. Butt, A.S., Haller, A., Xie, L.: A taxonomy of semantic web data retrieval techniques. In: International Conference on Knowledge Capture, pp. 1–9 (2015)
3. Drummond, N., Shearer, R.: The open world assumption. In: eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web, vol. 15, p. 1 (2006)
4. Gómez-Pérez, A.: Evaluation of ontologies. Int. J. Intell. Syst. **16**(3), 391–409 (2001)
5. Lee, B., Plaisant, C., Parr, C.S., Fekete, J.D., Henry, N.: Task taxonomy for graph visualization. In: AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (2006)
6. Rabitti, G., Khorrami Chokami, A., Coyle, P., Cohen, R.D.: A taxonomy of cyber risk taxonomies. Risk Anal. **45**(2), 376–386 (2025)
7. Senaratne, A.: Seka: Seeking knowledge graph anomalies. In: Companion Proceedings of the ACM Web Conference 2023, pp. 568–572 (2023)
8. Senaratne, A., Christen, P., Omran, P., Williams, G.: Anomaly detection and classification in knowledge graphs. arXiv preprint arXiv:2412.04780 (2024)
9. Senaratne, A., Christen, P., Williams, G., Omran, P.G.: Unsupervised identification of abnormal nodes and edges in graphs. ACM J. Data Inf. Qual. **15**(1), 1–37 (2022)
10. Senaratne, A., Omran, P.G., Christen, P., Williams, G.: TRIC: a triples corrupter for knowledge graphs. In: Pesquita, C., et al. (eds.) European Semantic Web Conference, pp. 117–122. Springer (2023). https://doi.org/10.1007/978-3-031-43458-7_22
11. Senaratne, A., Omran, P.G., Williams, G., Christen, P.: Unsupervised anomaly detection in knowledge graphs. In: Proceedings of the 10th International Joint Conference on Knowledge Graphs, pp. 161–165 (2021)
12. Sutcliffe, A., Rugg, G.: A taxonomy of error types for failure analysis and risk assessment. Int. J. Hum. Comput. Interact. **10**(4), 381–405 (1998)
13. Xue, B., Zou, L.: Knowledge graph quality management: A comprehensive survey. Trans. Knowl. Data Eng. **35**, 4969–4988 (2022)
14. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: a survey: a systematic literature review and conceptual framework. Semant. Web **7**(1), 63–93 (2015)