




Temporal Dimensions of Quality in Knowledge Graph Evolution: A Comprehensive Review

Yahia Atig^{1*}, Ahmed Zahaf², Nadri Khiati¹

¹ Department of Computer Science, University Center of Naâma, Naâma 45000, Algeria

² Department of Computer Science, University of Saïda, Saïda 20000, Algeria

Corresponding Author Email: y_atig@esi.dz

<https://doi.org/10.18280/isi.280428>

Received: 6 June 2023

Revised: 20 August 2023

Accepted: 29 August 2023

Available online: 31 August 2023

Keywords:

knowledge graph, data quality, Knowledge Graphs (KGs) evolution, temporal challenges, time representation

ABSTRACT

Efforts to enrich Knowledge Graphs (KGs) typically seek to augment data quality, semantic comprehension, and functional capabilities via the integration of various data sources. However, the inherent evolution of these sources over time potentially compromises the quality of the KGs. This paper provides a systematic exploration of the temporal challenges intrinsic to the progression of KGs, including the dynamics of changes, anomaly detection, the estimation of repair costs, and the delicate balance between changes and consistency. The complexities associated with the accurate representation of time in KGs are addressed, providing a critical assessment and understanding of this issue. A correction framework, bolstered by temporal considerations, is proposed, with an intent to scrutinize these techniques using various datasets in future research endeavors. This work represents a step forward in comprehending the quality of KGs by delving into their temporal aspects.

1. INTRODUCTION

The unprecedented surge in online data diversity, with respect to formats, types, and sources, has necessitated the development of a multitude of techniques and tools for data extraction, storage, processing, and analysis. The primary objective of these advances is the effective harnessing and management of the wealth of information inherent in this vast data landscape. A critical tool in this endeavor is the Knowledge Graph (KG). Despite not being a novel concept in its core—essentially resembling simplified versions of semantic networks [1]—KGs have recently garnered renewed interest due to their widespread industrial adoption. They serve as the backbone of key functionalities within intelligent virtual assistants such as Alexa, Siri, and Google Assistant. However, their significance extends beyond these applications, as KGs have demonstrated their utility across a myriad of domains, including artificial intelligence, linked data, big data, the Open Knowledge Network, and deep learning. This resurgence is synchronous with the advent of the Semantic Web and its associated research fields, underpinning the enduring relevance of KGs.

Like all interdisciplinary web technologies, KGs are subject to an array of emerging challenges. These challenges stimulate continual enhancements in the methods of KG creation and enrichment and necessitate the establishment of metrics for assessing the quality of the resultant structures. A particular issue that has surfaced in this context is the rectification of erroneous assertions within KGs. This paper delves into the solutions proposed in contemporary literature to address this problem, with a primary focus on Ontology Alignment techniques (a set of matchings) deployed between different KGs. Despite these proposals being in their nascent stages, this

analysis has instigated our exploration of a new issue termed the “Time Aspect in Knowledge Graph Evolution”. This aspect encapsulates the study of KG evolution in accordance with the shifts in knowledge within data sources, and concurrently, the temporal aspect of this evolution and the questions revolving around it.

The Time Aspect in Knowledge Graph Evolution acknowledges the dynamic nature of knowledge representation over time. It scrutinizes the impact of evolving data sources on the quality and accuracy of KGs. This aspect encompasses challenges such as discerning the various types of changes impinging on KGs, detecting anomalies that emerge during evolution, estimating the costs of repairing compromised information, and achieving a balance between incorporating minimal changes while preserving overall consistency. Moreover, the intricacies involved in accurately representing time within KGs are probed. Effective management of the Time Aspect is vital to ensure that KGs remain current, reliable, and accurate in reflecting the evolving nature of information.

The rest of this paper is structured as follows: Section II delineates various definitions that encapsulate the concept of KGs and reviews the history and applications of different forms of this notion, delineated into two eras: pre-2012 and post-2012. Section III outlines the potential methods employed in the creation and enrichment of KGs. Section IV discusses the quality of KGs, as determined by several metrics, while Section V elucidates the alignments in KGs and the process of assertion correction within the graph. Section VI defines the Time Aspect in Knowledge Graph Evolution, delineating three directions of prospects: KG evolution, Temporal KG, and Datasets for experimentation. Finally, a conclusion is provided in Section VII.

2. KNOWLEDGE GRAPH DEFINITIONS AND APPLICATIONS

To understand the importance of KGs in different computer systems, we first review a set of definitions and then its applications following the history of this data structure.

2.1 Definitions

Although there are a variety of Knowledge Graph definitions, in this section we quote the most famous and cited in the works dealing with KGs. The most concise definition is that proposed by Liu et al. [2]: “Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities”. For his part, Blumauer [3] has proposed: “Knowledge graphs could be envisaged as a network of all kind things which are relevant to a specific domain or to an organization. They are not limited to abstract concepts and relations but can also contain instances of things like documents and datasets”, while Paulheim [4] has divided KGs according to four points of view: “A knowledge graph (i) mainly describes real world entities and their interrelations, organized in a graph, (ii) defines possible classes and relations of entities in a schema, (iii) allows for potentially interrelating arbitrary entities with each other and (iv) covers various topical domains”. A last formal definition has been proposed by Färber et al. [5]: “We define a Knowledge Graph as an RDF graph. An RDF graph consists of a set of RDF triples where each RDF triple (s, p, o) is an ordered set of the following RDF terms: a subject $s \in U \cup B$, a predicate $p \in U$, and an object $U \cup B \cup L$. An RDF term is either a URI $u \in U$, a blank node $b \in B$, or a literal $l \in L$ ”.

Examination. In analyzing the definitions of KGs, several common threads emerge regarding their fundamental characteristics. A shared emphasis across all definitions is the integral role of entities and relationships within KGs. The studies [2, 4] go further by explicitly incorporating semantic types and properties into their conceptualizations, underscoring the importance of capturing nuanced attributes. Notably, Blumauer [3] introduces the notion of KGs being inherently tied to specific domains or organizational contexts, extending beyond abstract concepts. This aligns with Paulheim [4], who accentuates the breadth of KG applicability, spanning diverse topical domains. Additionally, the theme of schema and interrelations emerges, with advocating for a structured schema that delineates conceivable classes and relations of entities [4]. This notion is mirrored by Blumauer [3], who, along with Paulheim [4], acknowledges the potential for interconnecting arbitrary entities, thereby enabling a dynamic and versatile information landscape.

2.2 History and applications

The launch in 2012 of the Google Knowledge Graph led to dividing the lifespan of KGs into two eras: before 2012 and after that.

2.2.1 Knowledge graphs before 2012

The term KG was first used by Schneider [6] in the field of computerized education systems to represent and store an instruction course on a computer. At the same epoch, other works followed this form of representation using different names to implement their contributions. For instance, Kümmel [7] uses a KG in the linguistic field [8] in education to study

the impact of the knowledge units of an instructional course on teachers and students.

The 80s also saw a series of works on KGs. In a non-exhaustive way, we can cite the work of Rada [9] who defined a KG in the context of medical expert systems. Another KG was instantiated by Bakker [10] to represent cumulatively the contents of medical and sociological texts. In the following year, Rappaport and Gouyet [11] built a user interface for visualizing a knowledge base using a KG. Next, Srikanth and Jarke [12] instantiate a KG to modularize the knowledge required in software projects.

In the following decade, the notion of KG also saw the day in different contexts by several works. First, as a support to represent the collected knowledge from different experts [13]. Then, as a knowledge-based system that formally integrates knowledge from different sources for representing natural language by James [14]. And last but not least, as a Bayesian knowledge base that subsequently allows Bayesian inference to be applied by Shimony et al. [15].

From the 2000s, some works have improved the use of representation forms which are similar to KGs. An KG instance named “plan knowledge graph” was proposed by Jiang and Ma [16] to connect a set of goals using a set of dependencies. This makes it possible to find plans for particular purposes using search algorithms on this graph. In 2005, a KG for workflows was built by Helms and Buijsrogge [17] to represent the flow of knowledge in an organization. The actors of the workflow are represented by nodes while edges represent the flow of knowledge from one actor to another. Finally, Coursey and Mihalcea [18] proposed a dedicated KG to classify articles according to Wikipedia categories.

For what remains of the years preceding 2012, two relevant works have proposed representations as KGs. First, Pechsiri and Piriyaikul [19] used KG extraction from a text to graphically explain events (nodes) to users by revealing the causal relationships between these events (edges). And subsequently, Corby and Zucker [20] encode knowledge in KGs and provide the possibility to query such graphs.

Examination. Two discernible classes of KG applications come to the forefront. The first class, as demonstrated by Schneider [6], Rada [9], and Srikanth and Jarke [12], showcases KGs’ early utility in specific domains such as education, medicine, linguistics, and software projects. These instances underscore KGs’ role as specialized tools for representing and organizing knowledge within distinct contexts. The second class, encompassing the contributions of Jiang and Ma [16, 17, 19], unveils KGs’ evolving versatility as instruments for advanced information management. These works highlight KGs’ adaptability in facilitating diverse functions, such as plan formulation, workflow representation, and event-to-causality mappings. This classification of KG applications collectively illuminates the progression from domain-specific implementations to broader applications aimed at enhancing information organization and comprehension.

2.2.2 Knowledge graphs after 2012

The announcement of the Google Knowledge Graph in 2012 was primarily intended to improve the semantics of Google’s search function by making it possible to search for real-world objects (called things) instead of a simple match between strings. Although this announcement is informal and does not provide any implementation details [21], it has

boosted the adoption of the expression “Knowledge Graph” in the Semantic Web community (Following a search for the word “KG” made on January 01, 2023, we obtained about 5,910,000 results). Indeed, a family of famous applications have joined the wave of KGs since 2012. We present in the following list the most famous according to Wang et al. [22].

- **Freebase.** Launched in 2007 and closed in 2016 counting over 1.9 billion triplets (<https://developers.google.com/freebase/>), 4,000 types and 7,000 attributes. This open shared database was able to collect this considerable amount of data based on several web sources and also their users, such as Wikipedia (<https://www.wikipedia.org/>), NNDB (<https://www.nndb.com/>), Fashion Model Directory (<https://www.fashionmodeldirectory.com/>) and MusicBrainz (<https://www.wikidata.org/wiki/Wikidata>). After its closure, its data was transferred to Wikidata.
- **Wikidata** (<https://www.dbpedia.org/>). This project first appeared in 2012. With data now numbering over 55 million entities, Wikidata is based on a collection of websites, but primarily on sister Wikipedia projects, including Wikipedia, Wikivoyage, Wikisource and other websites.
- **DBpedia** (<https://wiki.dbpedia.org/about>). The most famous multilingual KG among this list and contains in its English versio (Following a search for the word "KG" made on January 01, 2023, we obtained about 5,910,000 results) more than 4.58 million entities and DBpedia data in 125 languages has 38.3 million entities. DBpedia data was extracted from Wikipedia and collaboratively edited by the community.
- **YAGO** (<https://datahub.io/collections/yago>). (Yet Another Great Ontology) is an open source KG that has more than 10 million entities and 120 million entity facts (Following a search for the word "KG" made on January 01, 2023, we obtained about 5,910,000 results). YAGO data comes mainly from Wikipedia, WordNet and GeoName.
- **ConceptNet** (<https://conceptnet.io/>). An open multilingual large-scale KG that contains over 8 million entities and 21 million entity links and originated from the Open Mind Common Sense project started in 1999.
- **CN-DBpedia** (<http://kw.fudan.edu.cn/>). A Chinese KG that has more than 16.8 million entities and 223 million entity links. Launched in 2017 For the purpose of extracting knowledge from Chinese encyclopedias.

Examination. The surge of KGs after 2012 underscores their growing scale, complexity, and practical utility. Google’s KG marked a shift toward nuanced understanding of real-world entities. Examples like Freebase, Wikidata, and CN-DBpedia highlight KGs’ capacity to handle vast data and accommodate diverse languages. Projects such as ConceptNet and YAGO showcase their ability to capture intricate relationships. This evolution signifies KGs’ pivotal role in organizing, connecting, and extracting insights from complex, diverse information sources.

The list presented above is a concise review involving only large-scale KGs. Indeed there is an indeterminate number of applications that emerged after 2012. A simple way to review these different types and sizes of KGs is to classify them according to a limited number of groups. Hogan et al. [23] have proposed four categories of KGs according to the objectives aimed by their owners:

- Category I: KGs that consist of nodes represent entities and edges represent relationships between those entities

(e.g., ConceptNet, CN-DBpedia).

- Category II: A KG in this class is a graph-structured knowledge base (e.g., Freebase, YAGO, Wikidata, DBpedia).
- Category III: Besides to the two past definitions, set of technical characteristics that a KG should comply with, such as: linking arbitrary entities and coverage of various topical areas [4]; acquiring and integrating information into an ontology and apply a reasoner to derive new knowledge [24], etc.
- Category IV: avoids providing a precise definition of KG and adopts an extensional definition, defining them by examples such as Google’s KG, Freebase, DBpedia and YAGO, among others [25] (e.g., Google KG, Freebase, DBpedia, YAGO).

3. KNOWLEDGE GRAPHS CREATION AND ENRICHMENT

It is important to discuss methodologies to create and then enrich KGs in order to fully understand it and its environment. Regarding the creation of a KG, an appropriate methodology depends on a set of factors, among them: the envisaged purposes and applications, the field, the actors involved, the available data sources, etc. These last can range from unstructured plain text to structured formats (including the whole range between the two). Such a process should be flexible and the result is an initial core which can be gradually enriched from other sources as needed. In this context, two examples are proposed in the KG community, namely: agile and “pay-as-you-go” methodologies.

(1) Agile methodology

according to Hunt and Thomas [26], the idea of keeping things simple is the guiding principle behind agile software development. This can be translated to the art of maximizing the amount of undone work to ensure simplicity. To do this and by taking steps as for instance constant refactoring, only essential features or framework functionality must be implemented in order to avoid future issues that may require additional work. Despite the admirable simplification offered by this principle, it is very common for developers to make one of the following two errors:

- Oversimplifying something that is really complex;
- Overcomplicating something that should be easy.

Example 1. Consider a team embarking on the creation of a KG to catalog medical research findings. Applying the agile methodology, the team prioritizes simplicity and incremental development. Initially, they focus on identifying essential entities like diseases, treatments, and researchers. Instead of attempting to define intricate relationships and attributes upfront, they opt for an “undone work” approach, concentrating on establishing a core framework. As they proceed, they constantly refactor the KG, iteratively adding complexity as needed. The team aims to avoid the mistake of oversimplification by ensuring that complex interconnections, such as multi-dimensional impacts of treatments, are not overlooked. Simultaneously, they guard against overcomplication by resisting the urge to introduce unnecessary layers of intricacy that could hinder future expansion. This agile strategy ensures that the KG’s foundation remains robust while allowing for gradual refinement as the project matures.

(2) “pay-as-you-go” methodology.

With the objective of overcoming the challenge of understanding complex database schemas which allows business users to quickly answer business questions and therefore Business Intelligence (BI) needs, Sequeda et al. [27] proposed a “pay-as-you-go” method based on the gradual construction of a KG allowing professional users to ask and answer their own questions with a minimum of IT support.

Example 2. Imagine a business intelligence team seeking to harness a Knowledge Graph to enhance decision-making. Adopting the “pay-as-you-go” methodology, they recognize the challenge of deciphering intricate database schemas. Rather than attempting a comprehensive upfront design, the team progressively constructs the KG to accommodate evolving business needs. They begin by creating a basic structure that captures high-level data categories, such as sales, customers, and products. This foundational structure enables immediate BI insights. As business users engage with the KG, they pose queries to the system, prompting the gradual expansion of the KG based on these inquiries. For instance, when a user seeks correlations between customer demographics and purchase patterns, the team adds relevant attributes and connections to the KG. This incremental growth empowers business users to interact with the KG autonomously, quickly obtaining insights without extensive IT involvement. By prioritizing gradual construction over a rigid schema, the “pay-as-you-go” approach adapts to evolving business demands while minimizing complexity.

Based on the work in Hogan et al. [23], the factors that influence the creation of a KG can cover on one side the human collaboration and on the other side the text, markup and structured sources.

3.1 Human collaboration

Like any system, human contribution (e.g., feedback, collaborative-editing platforms, crowd-sourcing platforms, etc.) can be very useful during the creation phase. However, despite some important KGs being primarily created through direct human efforts [28, 29], the involvement of humans, especially in extensive projects, introduces challenges such as high costs due to potential human errors and lack of consensus [30]. As a strategic response, an approach emerges that capitalizes on human expertise for nuanced tasks like validation [31] or preference assessment [32], while harnessing automation for streamlined data extraction.

Example 3. Consider a KG aiming to catalog biodiversity data, where automated algorithms efficiently populate species information while taxonomists validate and refine intricate classifications. This synergistic combination strikes a balance between precision and efficiency. To address cost-related concerns, stringent quality control measures involving expert review and collaborative platforms, along with clear guidelines and ontological frameworks, are implemented to ensure robustness and uniformity.

3.2 Text sources

Unstructured texts are valuable yet challenging data sources for KGs [33, 34]. Natural Language Processing (NLP) [35, 36] and Information Extraction (IE) [37-39] techniques unlock structured insights from these texts. NLP deciphers linguistic nuances, extracting entities and relationships.

Example 4. In medical literature, NLP identifies patient conditions and treatments. IE complements NLP by targeting

specific facts. Legal contracts can be scanned to capture parties and obligations, enhancing legal KGs. However, challenges like ambiguity and context variations persist. Algorithms must address these complexities for reliable extraction. Scalability and generalization also demand attention as data sources diversify, necessitating adaptable NLP and IE methodologies. NLP and IE enhance KGs by structuring unstructured data, though challenges require innovative solutions for accurate extraction and broader applicability.

3.3 Markup sources

Although not exclusive in its category (e.g., Wikitext for Wikipedia, TeX for typesetting, Markdown for Content Management Systems), HyperText Markup Language (HTML), as a World Wide Web Consortium (W3C) standard, predominates in web documents. While a rudimentary approach involves removing HTML tags to leave purified text for subsequent techniques, advanced extraction methods ([38, 40, 41]) leverage tags, exemplified by HTML elements like `<h1>`, RDFa annotations, XML tags, and microdata attributes. These markup tags, along with a spectrum of other languages, play a pivotal role in KG creation. XML, for instance, provides structured data annotation for domains like scientific research, while RDFa seamlessly integrates semantic annotations into web content. Markdown enhances textual annotations and collaborative documentation, contributing context to KGs. Specialized languages like MathML ensure accurate representation of mathematical concepts within mathematical KGs. In addition, a web browser in Corby and Faron-Zucker [42] can also produce an amalgamation of a local RDF graph with the DBpedia RDF graph, offering a consolidated perspective.

3.4 Structured sources

In the era of KGs, we can consider any previous structuring of data within or outside the web as a legacy form. The idea would therefore be to map this structured data to the KG under construction (or KG enrichment with more constraints). Compared to extraction techniques, such a mapping would be more beneficial in terms of data loss, since it actually allows the source data to be materialized in the form of a graph. This is an addition to the visualization advantages (creation of a graphical view on the legacy data).

4. KNOWLEDGE GRAPH QUALITY

The methodologies for creation and enrichment described earlier are reliant on the data available within the targeted sources. However, ensuring the completeness, consistency, and precision of this data, especially when originating from multiple sources, is not guaranteed. Thus, a vital undertaking involves evaluating the quality of the resultant knowledge graph. This evaluation process, too, faces variability in its definition contingent upon the objectives set forth during the initial creation and enrichment of the knowledge graph from external sources. Furthermore, the determination of quality aligns with the intended applications, purposes, domain, and contextual considerations.

Example 5. In a healthcare KG aggregating data from diverse medical databases, quality evaluation might emphasize data accuracy, timeliness, and adherence to medical

standards, reflecting the paramount significance of reliable clinical information in the KG.

Based on the dimensions cited in Batini et al. [43] which originate from the traditional domain of databases, Hogan et al. [23] introduced a set of quality dimensions tailored to the domain of KGs. We summarize these dimensions in the following points, providing examples grounded in the context of a healthcare KG:

4.1 Accuracy

An entity (or relation) encoded by a node (or edge) in a KG is considered accurate if it reflects a real phenomenon. Thereby, accuracy can be measured at three levels: (i) Syntactic accuracy which measures the degree of accuracy of the data relative to the grammatical rules defined for the domain and/or data model. (ii) Semantic accuracy which measures the rate of congruence between what is encoded and its image in the real world. And (iii) Timeliness accuracy which is a carbon copy of the previous dimension, but as a function of time: a knowledge graph can be semantically accurate now, but can quickly become inaccurate (obsolete) if no procedures are in place to keep it up to date in a timely manner.

Example 6. Based on Example 5, Syntactic Accuracy focuses on adhering to grammatical and structural rules for specific domains, as seen in a healthcare KG where precise medical terminology and formatting ensure accuracy. Semantic Accuracy is paramount in healthcare, ensuring KG data faithfully represents real-world medical phenomena to prevent inaccurate diagnoses and treatments. Timeliness Accuracy is vital due to the dynamic medical landscape, maintaining up-to-date information in the KG to avoid obsolete content and uphold its value for professionals and researchers.

4.2 Coverage

This dimension is used to avoid the loss of elements relevant to the domain during the creation or burial phase of the KG. Here too two sub-dimensions are present: (i) Completeness which measures the extent to which all the required information is present in a particular set of data. And (ii) Representativeness is used to assessing high-level biases in what is included/excluded from the knowledge graph [44].

Example 7. In the context of Example 5, the Coverage dimension becomes apparent when considering the inclusion of diverse medical data. Within Completeness, the KG should encompass a wide range of medical information, from patient profiles to treatment outcomes, to facilitate robust research. In terms of Representativeness, the KG should avoid biases by including a variety of medical conditions and demographics, ensuring balanced research outcomes and enhancing the KG's overall value for medical professionals and researchers.

4.3 Coherency

This measure aims to measure what is in conformity with respect to the formal semantics and the constraints defined at schema-level. It is divided into two sub-dimensions: (i) Consistency detects logical and/or formal contradictions. And (ii) Validity for detecting constraint violations, as captured by shape expressions [45].

Example 8. If we stay in the context of Example 5, the Coherency dimension ensures that medical data conforms to defined semantics and schema-level constraints. Consistency detects contradictions, such as conflicting medical information, while Validity identifies violations like inaccurate data formats or constraints, enhancing data accuracy and reliability.

4.4 Succinctness

This dimension to measure inclusion in a concise and intelligible way of what is only useful, and to avoid information overload.

Example 9. In the same context of Example 5, the Succinctness dimension focuses on presenting only pertinent and intelligible information, preventing information overload. It ensures that medical professionals and researchers can efficiently access essential medical insights without being overwhelmed by excessive or redundant data, thus optimizing the KG's usability and effectiveness.

5. CORRECTING ASSERTIONS AND ALIGNMENTS IN KNOWLEDGE GRAPHS

Efforts provided on the quality of the KGs can be reduced by several issues affecting their usability by different applications. Obvious examples of these issues are constraint violations and erroneous assertions. As part of the work on the quality of KGs, and in order to correct erroneous assertions (subject, predicate, object) and alignments (mappings) caused by lexical or semantic confusion (A very popular example in the English football context: The confusion between Manchester_United and Manchester_City, two football clubs based in Manchester, UK, can lead to facts about Manchester_United being incorrectly asserted about Manchester_City), Chen et al. [46] proposed a general correction framework based on the combination of four techniques: (i) Related Entity Estimation. to identify and classify, in decreasing order of similarity, the set of entities that are related to the correct object (substitute) of the assertion against a target assertion (to be corrected). (ii) Assertion Prediction used to estimate a likelihood score for each related entity of a target assertion to be corrected resulting from the first technique. (iii) Constraint-based Validation. where the authors use two kinds of soft constraints: property cardinality and hierarchical property range from the KG, then, they use a consistency checking algorithm to validate those candidate assertions. And (iv) Correction decision making. which acts in three stages: (a) Calculates the assertion likelihood score with an assertion prediction model, and the consistency score; (b) separately normalizes the two scores into [0; 1] according to all the predictions by the corresponding model; (c) ensembles the two scores by simple averaging them; and (d) filters out each candidate entity from the set of related entities if its calculated average is lower than a fixed threshold.

By combining the previously described techniques, the framework addresses specific challenges include:

- **Erroneous Assertions:** It focuses on correcting incorrect subject, predicate, and object assertions within the KG. It addresses cases where the KG contains inaccurate information due to lexical or semantic confusion.
- **Alignment Issues:** It also targets alignment problems, such as incorrect mappings between entities in the KG.

These issues can arise when entities are wrongly linked or mapped due to similarity or naming discrepancies.

- **Lexical and Semantic Confusion:** It is designed to handle issues caused by lexical ambiguity or semantic confusion, where entities with similar names or meanings are incorrectly associated in the KG.
- **Constraint Violations:** Soft constraints like property cardinality and hierarchical property range can lead to constraint violations within the KG. The framework includes a constraint-based validation step to ensure that candidate assertions adhere to these constraints.
- **Inconsistencies:** Inaccurate assertions can lead to inconsistencies in the KG. The framework employs a consistency checking algorithm to identify and address inconsistencies arising from incorrect assertions.

This framework presents a promising approach to addressing quality issues in KGs. However, it is not without limitations. One notable limitation is scalability. As the size of KGs increases, the computational complexity of techniques like related entity estimation and assertion prediction could pose challenges, potentially affecting the framework's performance and efficiency. Additionally, the framework's efficacy might diminish in cases of deep semantic complexity where nuanced contextual understanding is required for accurate entity substitution. Furthermore, the accuracy of the corrections heavily relies on the quality of the training data and initial assertions in the KG, making the framework sensitive to the overall data quality.

To enhance the framework's effectiveness, several potential improvements can be considered. (i) Integrating domain-specific knowledge and ontologies could bolster related entity estimation and assertion prediction, particularly in specialized fields where domain expertise is paramount. (ii) Exploring advanced machine learning techniques, such as deep learning architectures or ensemble methods, might elevate the accuracy of assertion prediction and correction decision making. (iii) Incorporating contextual information from surrounding entities and relations could empower the framework to better handle nuanced semantic contexts and disambiguate similar entities. Moreover, considering a feedback loop where human experts validate and contribute to the correction process could lead to continuous improvement. (iv) Addressing scalability challenges through parallel and distributed processing techniques could equip the framework to efficiently manage larger KGs without compromising performance.

Although the general framework proposed by Chen et al. [46] tries to deal with the problem of erroneous assertions as a whole. On the other hand, a set of works has focused on one

of the two obvious sub-problems; either detecting erroneous assertions or correcting them. In fact, most of the existing works are designed for detection and few of them offer corrective approaches. For instance, to check the consistency of KGs, techniques of assertions validation against logical constraints or rules are used in several works. Thereby, Topper et al. [47] employs a statistical analysis to enhance KGs with class disjointness, and property domain and range constraints, while Paulheim and Gangemi [48] enriched them via alignment with the DOLCE-Zero foundational ontology. Also, Kontokostas et al. [49] inspired from test-driven software engineering to provide an evaluation approach based on SPARQL query templates.

Correction methods can be divided according to whether they are KG-specific or not. In the first set, Pellissier Tanon et al. [50] target constraint violations in Wikidata and built their correction rules based on the editing history of this knowledge base. Besides, Dimou et al. [51] convert Wikipedia relational data to DBpedia knowledge mappings to fix errors during DBpedia construction. Obviously, these techniques are dependent on additional meta-information from the knowledge base. In contrast, a second set of more general approaches (KG-independent), aims to eliminate constraint violations. For example, Lertvittayakumjorn et al. [52] correct assertions by substituting objects or subjects with correct entities obtained by keyword matching. Similarly, Melo and Paulheim [53] replace erroneous entities using a collection of replacement candidates from Wikipedia disambiguation page (if exists) and then their classification by lexical similarity. Furthermore, Chortis and Flouris [54] anticipate the violation of integrity constraints by injecting new properties into the base. This ensures the consistency of the KG but fails to correct erroneous assertions that satisfy the constraints.

Correction methods in KGs exhibit a trade-off between KG-specific and KG-independent approaches. KG-specific methods, like those by Pellissier Tanon et al. [50, 51], offer high accuracy tailored to the KG's structure but may lack generalizability. KG-independent methods, exemplified by Lertvittayakumjorn et al. [52, 53], provide broader applicability across KGs but might sacrifice some accuracy due to their detachment from KG-specific nuances. Balancing accuracy and generality guides the choice between these approaches based on the KG's characteristics and desired outcomes.

6. TIME ASPECT IN KNOWLEDGE GRAPH EVOLUTION

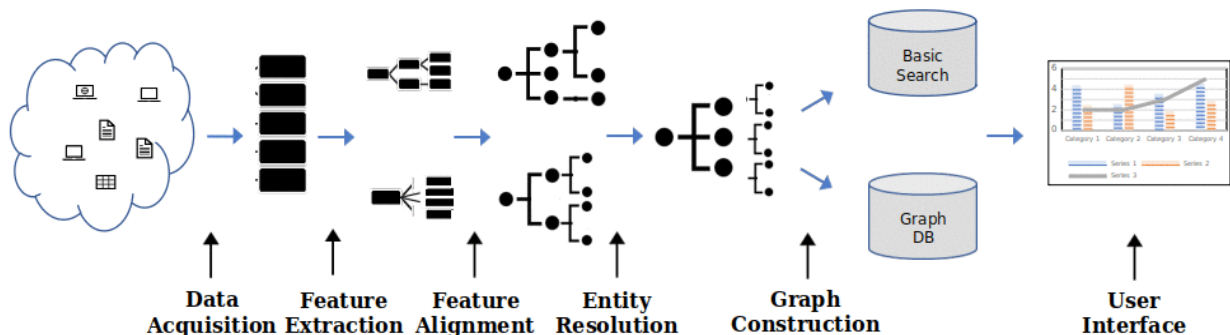


Figure 1. KG construction process [55]

Day after day the community of KGs faces new challenges, as an example we can cite the techniques that revolve around the four dimensions of KG quality discussed in Section 4 (accuracy, coverage, coherency and succinctness), but not only. According to Baclawski et al. [55], challenges exist at every stage of the process for building a KG. This process, as illustrated in Figure 1, transforms unprocessed, disparate data into a more refined, structured, and interconnected product. The outcome is a dataset that is not only easier to query, analyze, and visualize, but also well-organized and cleansed, enhancing its ease of interrogation and analysis.

Inspired by the points discussed in Section 4 (especially timeliness accuracy and coherency) together with the open issues and questions discussed in Baclawski et al. [55], it is possible to define a new research track according to a view that groups the quality of KGs together with the notion of time. We consider the two axes to be important to cover.

6.1 Knowledge graph evolution

Given the importance of KGs cited throughout this survey, and given that the process of construction and enrichment of KGs is a very difficult and complicated process (see Section 3 and Figure 1), which is based on knowledge conveyed by different sources (structured, semi-structured or not), and these sources are continually evolving so that they reflect our progressive understanding of reality [56], it is very logical, at least theoretically, that the evolution of knowledge in data sources can affect the quality of KGs. Therefore, we consider it important to answer the following questions:

(1) **Change impact.** What type of change within the data sources affects the KG? What is the impact of each change [57]? Are there any changes that can be ignored at the KG level? Do we need to put constraints on the changes? Is there any impact of alignment quality principles such as the conservativity of alignment [58] on KGs?

(2) **Anomalies Detection.** What kind of anomaly is considered to be a violation affecting the quality of the KG? Which method is adequate for the detection of each type of violation? What are the possibilities of reducing or adapting this problem to other mature problems in order to exploit the already existing infrastructure?

(3) **Repair cost.** What is the estimated cost to repair the violations in the KG? At what point repair will be unnecessary and reconstruction is required? Do we adopt an adaptive approach or a compute from scratch? What is the balance between the principle of minimal change is the consistency in the KGs?

6.2 Temporal knowledge graph

The problem of representing time is one of the challenges faced by KGs [55]. This notion is difficult to manage to the point where the majority of IA systems (even the most powerful) may encounter challenges in effectively handling temporal aspects. To clarify this, we adapt an example presented in the study of Baclawski et al. [55].

Example 10. If you ask Google “How old is Manchester-united?” or “How old is Manchester-city?”, you get the correct answers; but if you ask “Who is older, Manchester-united or Manchester-city?”, all you get are links to articles that mention both football clubs. This problem did not arise from the lack of temporal entities in the KGs, but rather because they did not exactly reflect their nature in the real world. In this context, a

relevant question, first raised in study [59], interests us is the following:

Temporal aspect. What are the abstract and concrete aspects of time that the KG research community should explore? What are the aspects of general reasoning and specific applications? What are the advantages of temporal reasoning for AI systems?

In addressing these issues, we also plan to improve the general framework in Chen et al. [46] explored in the Section 5 in the following points:

(1) How to involve the notion of time in the similarity between the related assertions and the target assertion.

(2) If we adopt an adaptive approach, how to better exploit the Sub-graph extraction technique in the Assertion Prediction step?

(3) How to improve the consistency checking algorithm so that it iterates only according to the parts dependent on the changes?

6.3 Dataset for experimentation

One of the major hindrances to validate any study results is the lack of datasets, especially for previously unreleased issues. To lessen this task for future contributions, we suggest an evaluation of the execution time between such proposals that may emerge in this context and the original general framework in Chen et al. [46]. For instance, this experiment can be carried out concurrently between the two frameworks on the same dataset to observe the usefulness of an adaptive approach compared to the compute from scratch [60]. A suitable dataset can be the same one proposed in Chen et al. [46], with a targeted modification to extract new modified KGs versions. Or it can be represented by multiple versions of KG built from a frequently changing data source (e.g., Dbpedia (<https://wiki.dbpedia.org/>)).

7. CONCLUSION

The objective of this survey was to draw up a clear and precise definition of a new problem called “Time Aspect in Knowledge Graph Evolution”. To do this, we analyzed the evolution over successive decades of different points of view on KGs and their applications. Then, we discussed a set of methodologies to create and enrich KGs followed by four metrics to measure their qualities. Subsequently, we presented what has been done as techniques to correct the erroneous assertions which allowed to bring to light the time aspect during the KG evolution through three sets of open issues: KG evolution, Temporal KG and Datasets for experimentation.

REFERENCES

- [1] Lehmann, F. (1992). Semantic networks. *Computers & Mathematics with Applications*, 23(2-5): 1-50. [https://doi.org/10.1016/0898-1221\(92\)90135-5](https://doi.org/10.1016/0898-1221(92)90135-5)
- [2] Liu, S., d’Aquin, M., Motta, E. (2017). Measuring accuracy of triples in knowledge graphs. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland*, pp. 343-357. https://doi.org/10.1007/978-3-319-59888-8_29
- [3] Blumauer, A. (2014) From Taxonomies over Ontologies to Knowledge Graphs. Retrieved on February 2023 from:

- <https://blog.semantic-web.at/2014/07/15/from-taxonomies-over-ontologies-to-knowledgegraphs>.
- [4] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3): 489-508. <https://doi.org/10.3233/SW-160218>
- [5] Färber, M., Bartscherer, F., Menne, C., Rettinger, A. (2018). Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1): 77-129. <https://doi.org/10.3233/SW-170275>
- [6] Schneider, E.W. (1973). Course modularization applied: the interface system and its implications for sequence control and data analysis. In *Association for the Development of Instructional Systems (ADIS)*, Chicago, Illinois, pp. 1-17.
- [7] Kümmel, P. (1980). An algorithm of limited syntax based on language universals. *Computational and mathematical linguistics: Proceedings of the International Conference on Computational Linguistics*: Pisa, Pisa, Romanicum, pp. 225-247. <https://doi.org/10.1400/274144>
- [8] Marchi, E., Miguel, O. (1974). On the structure of the teaching-learning interactive process. *International Journal of Game Theory*, 3(2): 83-99. <https://doi.org/10.1007/BF01766394>
- [9] Rada, R. (1986). Gradualness eases refinement of medical knowledge. *Medical Informatics*, 11(1): 59-73. <https://doi.org/10.3109/14639238608994975>
- [10] Bakker, R.R. (1987). Knowledge Graphs: Representation and structuring of scientific knowledge. Ph.D. Dissertation. University of Twente.
- [11] Rappaport, A.T., Gouyet, A.M. (1988). Dynamic, interactive display system for a knowledge base. US Patent US4752889A. Washington, DC: U.S. Patent and Trademark Office.
- [12] Srikanth, R., Jarke, M. (1989). The design of knowledge-based systems for managing ill-structured software projects. *Decision Support Systems*, 5(4): 425-447. [https://doi.org/10.1016/0167-9236\(89\)90020-1](https://doi.org/10.1016/0167-9236(89)90020-1)
- [13] Dieng, R., Giboin, A., Tourtier, P.A., Corby, O. (1992). Knowledge acquisition for explainable, multi-expert, knowledge-based design systems. In *Current Developments in Knowledge Acquisition-EKAW'92: 6th European Knowledge Acquisition Workshop Heidelberg and Kaiserslautern*, Springer Berlin Heidelberg, 298-317. https://doi.org/10.1007/3-540-55546-3_47
- [14] James, P. (1992). Knowledge graphs. In *1991 Workshop on Linguistic Instruments in Knowledge Engineering*, Elsevier, 97-117.
- [15] Shimony, S.E., Domshlak, C., Santos, E. (1997). Cost-sharing in bayesian knowledge bases. In *UAI'97: Proceedings of the Thirteenth conference on Uncertainty in Artificial Intelligence*, 421-428.
- [16] Jiang, Y.F., Ma, N. (2002). A plan recognition algorithm based on plan knowledge graph. *Journal of Software*, 13(4): 686-692.
- [17] Helms, R., Buijsrogge, K. (2005). Knowledge network analysis: A technique to analyze knowledge management bottlenecks in organizations. In *16th International Workshop on Database and Expert Systems Applications (DEXA'05)*, Copenhagen, Denmark, pp. 410-414. <https://doi.org/10.1109/DEXA.2005.127>
- [18] Coursey, K., Mihalcea, R. (2009). Topic identification using wikipedia graph centrality. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Companion, Boulder Colorado pp. 117-120. <https://doi.org/10.3115/1620853.1620887>
- [19] Pechsiri, C., Piriyaikul, R. (2010). Explanation knowledge graph construction through causality extraction from texts. *Journal of Computer Science and Technology*, 25(5): 1055-1070. <https://doi.org/10.1007/s11390-010-9387-0>
- [20] Corby, O., Zucker, C.F. (2010). The KGRAM abstract machine for knowledge graph querying. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, Canada, 338-341. <https://doi.org/10.1109/WI-IAT.2010.144>
- [21] Bergman, M.K. (2019). A common sense view of knowledge graphs. *Adaptive Information, Adaptive Innovation, Adaptive Infrastructure Blog*.
- [22] Wang, P.L., Jiang, H., Xu, J.F., Zhang, Q. (2019). Knowledge graph construction and applications for web search and beyond. *Data Intelligence*, 1(4): 333-349. https://doi.org/10.1162/dint_a_00019
- [23] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., Ngomo, A.N., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4): 1-37. <https://doi.org/10.1145/3447772>
- [24] Ehrlinger, L., Wöß, W. (2016). Towards a definition of knowledge graphs. In *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems-SEMANTiCS2016 Leipzig, Germany*. <http://ceur-ws.org/Vol-1695/paper4.pdf>.
- [25] Bonatti, P.A., Decker, S., Polleres, A., Presutti, V. (2019). Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371). In *Dagstuhl Reports*, 8(9): 29-111. <https://doi.org/10.4230/DagRep.8.9.29>
- [26] Hunt, A., Thomas, D. (2003). The trip-packing dilemma [agile software development]. *IEEE Software*, 20(3): 106-107. <https://doi.org/10.1109/MS.2003.1196331>
- [27] Sequeda, J.F., Briggs, W.J., Miranker, D.P., Heideman, W.P. (2019). A pay-as-you-go methodology to design and build enterprise knowledge graphs from relational databases. In *The Semantic Web-ISWC 2019: 18th International Semantic Web Conference*, Springer International Publishing, Auckland, New Zealand, pp. 526-545. https://doi.org/10.1007/978-3-030-30796-7_32
- [28] He, Q., Chen, B.C., Agarwal, D. (2016). Building the linkedin knowledge graph. *Linkedin Blog*. <https://engineering.linkedin.com/blog/2016/10/building-the-linkedin-knowledge-graph>.
- [29] Vrandečić, D., Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78-85. <https://doi.org/10.1145/2629489>
- [30] Paulheim, H. (2018). How much is a Triple? Estimating the Cost of Knowledge Graph Creation, *CEUR Workshop Proceedings*, Sun SITE Central Europe (CEUR), 4. http://ceur-ws.org/Vol-2180/ISWC_2018_Outrageous_Ideas_paper_10.pdf.
- [31] Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., Pintscher, L. (2016). From freebase to wikidata: the

- great migration. In Proceedings of the 25th International Conference on World Wide Web, pp. 1419-1428. <https://doi.org/10.1145/2872427.2874809>
- [32] Jurgens, D., Navigli, R. (2014). It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2: 449-464. https://doi.org/10.1162/tacl_a_00195
- [33] Hellmann, S., Lehmann, J., Auer, S., Brümmer, M. (2013). Integrating NLP using linked data. In *The Semantic Web-ISWC 2013: 12th International Semantic Web Conference*, Sydney, Springer Berlin Heidelberg, 98-113. https://doi.org/10.1007/978-3-642-41338-4_7
- [34] Rospocher, M., Van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37: 132-151. <https://doi.org/10.1016/j.websem.2015.12.004>
- [35] Jurafsky, D., Martin, J.H. (2019). Speech and language processing. https://web.stanford.edu/~jurafsky/slp3/old_oct19/17.pdf
- [36] Maynard, D., Bontcheva, K., Augenstein, I. (2017). Natural language processing for the semantic web. Springer Cham. <https://doi.org/10.1007/978-3-031-79474-2>
- [37] Grishman, R. (2012). Information extraction: Capabilities and challenges. Technical Report. NYU Dept. CS. International Winter School in Language and Speech Technologies.
- [38] Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I. (2020). Information extraction meets the semantic web: A survey. *Semantic Web*, 11(2): 255-335. <https://doi.org/10.3233/SW-180333>
- [39] Weikum, G., Theobald, M. (2010). From information to knowledge: Harvesting entities and relationships from web sources. In Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 65-76. <https://doi.org/10.1145/1807085.1807097>
- [40] Lockard, C., Dong, X.L., Einolghozati, A., Shiralkar, P. (2018). CERES: Distantly supervised relation extraction from the semi-structured web. arXiv Preprint arXiv: 1804.04635. <https://doi.org/10.48550/arXiv.1804.04635>
- [41] Lu, C.L., Bing, L.D., Lam, W., Chan, K., Gu, Y. (2013). Web entity detection for semi-structured text data records with unlabeled data. *International Journal of Computational Linguistics and Applications*, 4(2): 135-150.
- [42] Corby, O., Faron-Zucker, C. (2016). Graph transformation language and server for the Web of data. *Revue d'Intelligence Artificielle*, 30(5): 607-627. <https://doi.org/10.3166/RIA.30.607-627>
- [43] Batini, C., Rula, A., Scannapieco, M., Viscusi, G. (2015). From data quality to big data quality. *Journal of Database Management (JDM)*, 26(1): 60-82. <https://doi.org/10.4018/JDM.2015010103>
- [44] Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, 61(6): 54-61. <https://doi.org/10.1145/3209581>
- [45] Thornton, K., Solbrig, H., Stupp, G.S., Labra Gayo, J.E., Mietchen, D., Prud'Hommeaux, E., Waagmeester, A. (2019). Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation. In *The Semantic Web: 16th International Conference*, Springer International Publishing, pp. 606-620. https://doi.org/10.1007/978-3-030-21348-0_39
- [46] Chen, J.Y., Jiménez-Ruiz, E., Horrocks, I., Chen, X., Myklebust, E.B. (2021). Correcting assertions and alignments of large scale knowledge bases. In *Semantic Web Journal*, 14(3): 1-25. <https://doi.org/10.3233/SW-210448>
- [47] Töpfer, G., Knuth, M., Sack, H. (2012). DBpedia ontology enrichment for inconsistency detection. In Proceedings of the 8th International Conference on Semantic Systems, pp. 33-40. <https://doi.org/10.1145/2362499.2362505>
- [48] Paulheim, H., Gangemi, A. (2015). Serving DBpedia with DOLCE-more than just adding a cherry on top. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference*, Springer International Publishing, Bethlehem, PA, USA, pp. 180-196. https://doi.org/10.1007/978-3-319-25007-6_11
- [49] Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A. (2014). Test-driven evaluation of linked data quality. In Proceedings of the 23rd International Conference on World Wide Web, pp. 747-758. <https://doi.org/10.1145/2566486.2568002>
- [50] Pellissier Tanon, T., Bourgaux, C., Suchanek, F. (2019). Learning how to correct a knowledge base from the edit history. In *The World Wide Web Conference*, pp. 1465-1475. <https://doi.org/10.1145/3308558.3313584>
- [51] Dimou, A., Kontokostas, D., Freudenberg, M., Verborgh, R., Lehmann, J., Mannens, E., Hellmann, S., Van de Walle, R. (2015). Assessing and refining mappingsto rdf to improve dataset quality. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference*, Springer International Publishing, 133-149. https://doi.org/10.1007/978-3-319-25010-6_8
- [52] Lertvittayakumjorn, P., Kertkeidkachorn, N., Ichise, R. (2017). Correcting range violation errors in DBpedia. In *International Semantic Web Conference Posters, Demos & Industry Tracks*.
- [53] Melo, A., Paulheim, H. (2017). An approach to correction of erroneous links in knowledge graphs. In *CEUR Workshop Proceedings*, RWTH Aachen, 2065: 54-57.
- [54] Chortis, M., Flouris, G. (2015). A diagnosis and repair framework for DL-LiteA KBs. In *European Semantic Web Conference*, Cham: Springer International Publishing, pp. 199-214. https://doi.org/10.1007/978-3-319-25639-9_37
- [55] Baclawski, K., Bennett, M., Berg-Cross, G., Schneider, T., Sharma, R., Singer, J., Sriram, R.D. (2021). Ontology summit 2020 communiqué: Knowledge graphs. *Applied Ontology*, 16(2): 229-247. <https://doi.org/10.3233/AO-210249>
- [56] Hepp, M. (2008). Ontologies: State of the art, business potential, and grand challenges. *Ontology Management: Semantic Web, Semantic Web Services, and Business Applications*, 3-22. https://doi.org/10.1007/978-0-387-69900-4_1
- [57] Atig, Y., Amine, A., Zahaf, A., Senthil Kumar, A.V. (2013). Alignment evolution between ontologies using a change log. *International Journal of Data Mining and Emerging Technologies*, 3(2): 81-87. <https://doi.org/10.5958/j.2249-3220.3.2.011>

[58] Atig, Y., Zahaf, A., Bouchiha, D. (2016). Conservativity principle violations for ontology alignment: Survey and trends. *International Journal of Information Technology and Computer Science (IJITCS)*, 8(7): 61-71. <https://doi.org/10.5815/ijitcs.2016.07.09>

[59] Davis, E. (2020) Time and space in knowledge graphs.

Retrieved on February 2023 from <https://go.aws/2SxOVZ9>.

[60] Atig, Y., Zahaf, A., Bouchiha, D., Malki, M. (2022). Alignment conservativity under the ontology change. *Journal of Information Technology Research (JITR)*, 15(1): 1-19. <https://doi.org/10.4018/JITR.299923>