Uncertainty Management in the Construction of Knowledge Graphs: a Survey

Miguel Couceiro, Lucas Jarnac, Yoan Chabot

Uncertainty Management in the Construction of Knowledge Graphs: a Survey

Lucas Jarnac [a,b,*], Yoan Chabot [a] and Miguel Couceiro [b,c]
[a] *Orange, France*
E-mails: lucas.jarnac@orange.com, yoan.chabot@orange.com
[b] *Université de Lorraine, CNRS, LORIA, Nancy, France*
E-mail: miguel.couceiro@loria.fr
[c] *INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal*

Abstract. Knowledge Graphs (KGs) are a major asset for companies thanks to their great flexibility in data representation and their numerous applications, *e.g.,* vocabulary sharing, Q/A or recommendation systems. To build a KG it is a common practice to rely on automatic methods for extracting knowledge from various heterogeneous sources. But in a noisy and uncertain world, knowledge may not be reliable and conflicts between data sources may occur. Integrating unreliable data would directly impact the use of the KG, therefore such conflicts must be resolved. This could be done manually by selecting the best data to integrate. This first approach is highly accurate, but costly and time-consuming. That is why recent efforts focus on automatic approaches, which represents a challenging task since it requires handling the uncertainty of extracted knowledge throughout its integration into the KG. We survey state-of-the-art approaches in this direction and present constructions of both open and enterprise KGs and how their quality is maintained. We then describe different knowledge extraction methods, introducing additional uncertainty. We also discuss downstream tasks after knowledge acquisition, including KG completion using embedding models, knowledge alignment, and knowledge fusion in order to address the problem of knowledge uncertainty in KG construction. We conclude with a discussion on the remaining challenges and perspectives when constructing a KG taking into account uncertainty.

Keywords: Knowledge reconciliation, Uncertainty, Heterogeneous sources, Knowledge graph construction

1. Introduction

Huge amounts of data expressed in the form of tables, texts, or databases are generated by organizations every day. When using these data within an organization, we have to deal with uncertainty, as the data often suffer from contradictions and differences in specificity. These are the effect of incompleteness, vagueness, fuzziness, invalidity, ambiguity, and timeliness leading to uncertainty about the correctness of the data [28]. The uncertainty can be due to the source of the data (*e.g.,* a document written by an expert *vs.* a non-expert in the field concerned) or in the data itself (*e.g.,* a scientific supposition where the fact is not yet well-defined but accepted by consensus as it is). For example, on the French Wikipedia page of the former president of France Jacques Chirac[1], we can read that he was the mayor of Paris from March 25, 1977 to May 16, 1995 while on Wikidata[2] it is mentioned that he was the mayor of Paris from March 20, 1977 to May 16, 1995 as depicted in Figure 1. In addition, data are not settled over time [118]. Some facts are known to change including all facts that involve a period of time for which the fact is

*Corresponding author. E-mail: lucas.jarnac@orange.com.
[1] https://fr.wikipedia.org/wiki/Jacques_Chirac
[2] https://www.wikidata.org/wiki/Q2105

(a) French Wikipedia page of Jacques Chirac.

(b) Wikidata page of Jacques Chirac.

Fig. 1. Illustration of a contradiction between two sources. One of the sources claims that the mandate of Jacques Chirac as mayor of Paris began on March 25 while the other claims that it began on March 20.

valid (*e.g.,* the mandate of a president or the place of residence of a person) or knowledge in specific domains may be known to change regularly, which is often the case for History where excavations can modify knowledge which, until today, was considered established.

Many knowledge graphs (KGs) have been built to represent such data in recent years, and they have become major asset for organizations, since KGs can support various downstream tasks such as knowledge and vocabulary indexation, as well as other applications in recommendation systems, question/answering systems, knowledge management, or search engine systems [61, 109]. To build or enrich a KG and reconcile uncertain data, we can rely on manual approaches (*e.g.,* domain experts) but this is a time-consuming and tedious process. Alternatively, it is common to leverage automatic knowledge extraction approaches that handle large volumes of data from various heterogeneous sources, *e.g.,* texts [35], tables [90], or databases to ensure the coverage of the KG. These automatic approaches are usually based on three main steps:

1. Extraction of knowledge from documents.
2. Detection of duplicates and conflicts between extracted knowledge. Conflicts occur from differences of specificity or knowledge contradictions.
3. Fusion of aligned knowledge: once detection is completed, conflicting knowledge should be reconciled.

However, each of the aforementioned steps is error-prone and increases the uncertainty on extracted knowledge due to the performance of the algorithms [72, 83, 146].

Uncertainty may also be found in knowledge, which we can distinguish two types: *objective knowledge* in which a single value is accepted (*e.g.,* the mandate of Jacques Chirac where only one period of time is the true value), and *subjective knowledge* in which several values can be accepted according to their context and point of view (*e.g.,* the number of participants in a protest depending on the counting technique). Most KG construction methods do not take into account noisy facts and the uncertainty inherent in extraction algorithms and knowledge, which may impact downstream applications. Therefore, there is a need to reconcile knowledge units extracted from heterogeneous sources before integrating them into the KG in order to obtain a single or multiple representations that are as reliable and fair as possible [109]. In this survey, we review different approaches to integrate knowledge uncertainty in the main steps of KG construction with up-to-date knowledge fusion methods and its representation in the graph. [114] surveys approaches and evaluation methods for KG refinement, particularly KG completion and error detection methods. In [87], the authors review truth discovery methods used in knowledge fusion before 2015. However, to the best of our knowledge, there is no survey about uncertainty handling in the construction of KGs.

The remainder of this survey is structured as follows. In Section 2, we describe our research methodology which led us to write this survey. We introduce the definition of KGs with some well-known KGs that have been built in recent years, then tools and quality metrics considered in KGs construction in Section 3. We present some knowledge extraction approaches and why they lead to uncertain knowledge in Section 4. An ideal knowledge integration pipeline handling the uncertainty of knowledge is provided in Section 5, while the steps of knowledge refinement from the pipeline are described in Section 6 for uncertainty consideration in KG representation learning, in Section 7 for knowledge alignment, and in Section 8 for knowledge fusion. The solutions for uncertainty representation in KGs are then listed and depicted in Section 9. We also discuss some perspectives on the use of uncertainty in the KG ecosystem in Section 10, before concluding this survey in Section 11.

2. Research Methodology

This paper aims at surveying methods to construct a KG from uncertain knowledge. In this section, we provide our research methodology for discovering and selecting papers on this purpose. To find papers of interest, we mainly used the Google Scholar search engine and created alerts with the following keywords: KG fusion, multi-source knowledge fusion, KG resolution, KG quality, knowledge fusion, KG reconciliation, KG alignment, KG matching, KG resolution, KG cleaning. The aforementioned keywords have been combined with the keyword "uncertain" and the terms "knowledge", "data", and "information" used interchangeably, *e.g.,* "uncertain knowledge fusion", "uncertain data fusion", or "uncertain information fusion".

For the selection of the papers, we proceeded as follows: (1) we look at the title of the paper, if it is relevant and seems to be related to one of the topics we were looking for, then we read the abstract; (2) if the paper presents a method related to uncertain KG construction or a method that is not related to KGs but which could be extrapolated to a KG, we selected it. We provide Figure 2 that depicts the distribution of paper publication years according to four applications of KG refinement: uncertain embedding, knowledge fusion, knowledge alignment, and uncertainty representation. We observe that the representation of uncertainty has been addressed in the literature since 2004, while the vector space representation of uncertainty is a more recent research topic (since 2016). The number of papers on knowledge fusion mentioned in this survey has remained constant over the last few years. This is due to new models based on deep learning that are being explored to tackle these tasks.

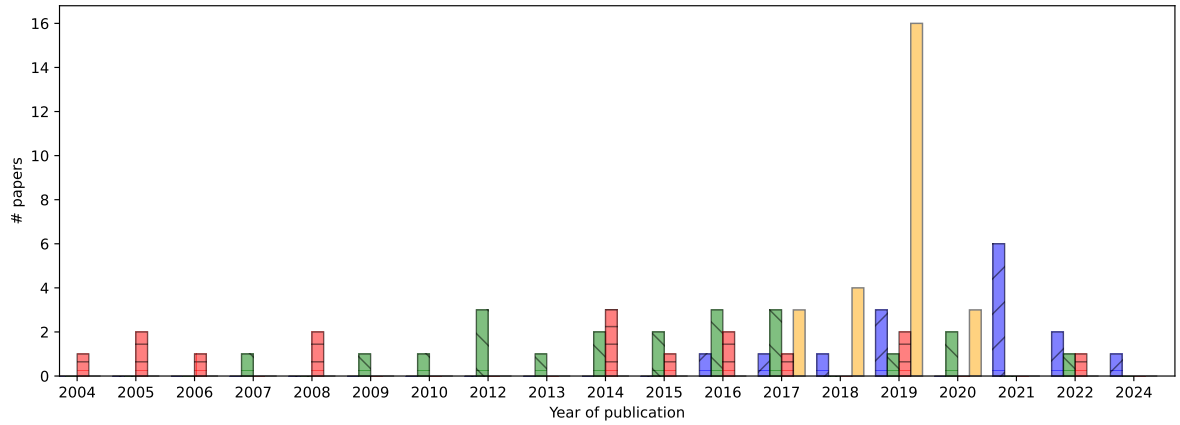

Fig. 2. Distribution of paper publication years according to uncertain KG embedding ( ), knowledge fusion ( ), knowledge alignment ( ), uncertainty representation ( ).

3. Knowledge Graphs

Before going into further detail on reconciliation approaches, it is important to define KGs, which are the core of this survey. KGs provide a structural representation of knowledge that is captured by the relations between entities in the graph. The KGs provide a concise and intuitive data representation and abstraction, making them an ideal tool to manage knowledge of organizations in a sustainable way or to support search and querying applications [60], which led several companies to build their own KGs [109].

In this section, we provide a definition of KG, and we describe some well-known KGs including open KGs and Enterprise Knowledge Graphs (EKGs) and how their consistency is maintained in Section 3.1.

3.1. What is a Knowledge Graph?

What is the meaning of knowledge? Data are uninterpretable signals (*e.g.,* numbers or characters). Information is data equipped with a meaning. In [122], the authors define knowledge as data and information that enter into a generative process supporting tasks and creating new information. A knowledge graph (KG) is "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities" [60].

Formally, KGs are directed and labeled multigraphs $(\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations, and $\mathcal{T}$ is the set of triples $\langle$subject, predicate, object$\rangle$ that are the atomic elements of KGs, also called facts. The subject and object are represented by nodes, the predicate indicates the nature of the relationship holding between the subject and object represented by an edge in the KG [60, 109]. For instance, such a triple could be $\langle$Suwon-si, location, Korea$\rangle$ as illustrated in Figure 3. The classes and relationships of entities are defined through a schema or otherwise named an ontology, which can be itself represented as a graph embedded in the KG [30, 36].

In a KG, two boxes coexist, namely Terminology Box (TBox) and Assertion Box (ABox). TBox defines classes and properties and ABox contains instances of classes defined in the TBox. For example, in Figure 3, "Company" and "Country" are concepts defined by the ontology, "Galaxy S23", "Samsung", and "Korea" are the instances of these concepts while "2023", "800€" are literals *i.e.,* attributes that characterize an entity. This data representation in semi-structured form, defined by its ontology, offers a clear and flexible semantic representation whose classes and relations can easily be added and connect a large number of domains [30].

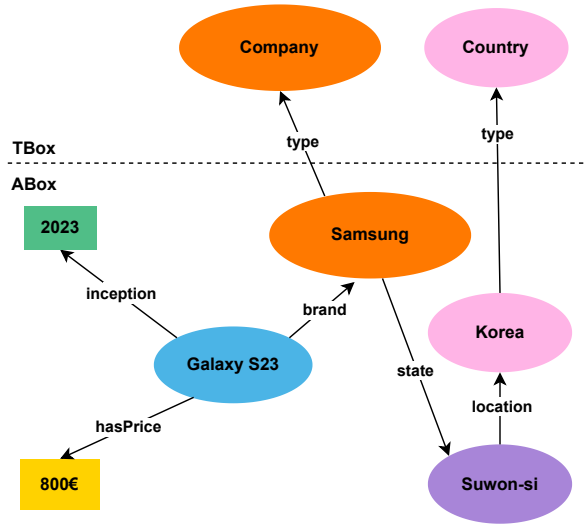

Fig. 3. TBox stands for terminology box that contains classes and properties; ABox stands for assertion box that contains instances (*e.g.,* Galaxy S23) and values (*e.g.,* "2023").

3.2. Open Knowledge Graphs

For the last few years, some KG construction projects that link general knowledge about the world have appeared. The best known is probably **Wikidata**[3], a large, free, and collaborative KG supported by the Wikimedia Foundation. It is a multilingual general-purpose KG that contains more than 100 million elements [138]. The structure of Wikidata is based on property-value pairs, where each property and entity is an element. A typical entity also contains labels, aliases, descriptions, links and mentions to Wikipedia articles. To maintain the quality of data, some constraints such as properties or unique values constraints alert the user in a case of suspicion of the input content

[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

(*e.g.,* constraint violations). Wikidata also keeps the sources and references of provenance of entities to ensure their traceability, which is one of the requirements for KG quality [145].

NELL is a research project led by Carnegie Mellon University in which an intelligent computer agent that ran continuously between January 2010 and September 2018 according to the official NELL website[4] and that every day extracted knowledge from texts, tables, and lists from the web to feed a Knowledge Base (KB) [16]. To maintain the consistency of the KB, the knowledge integrator of NELL exploits relationships between predicates by respecting mutual exclusion and type checking information. On top of that, NELL components provide a probability for each candidate and a summary of the source supporting it, hence which can be qualified as a probabilistic KB.

YAGO is an ontology built on statements of Wikipedia that combines high coverage with high quality [128]. The data model of YAGO is based on entities and binary relations extracted from WordNet and Wikipedia. A manual evaluation is performed to verify the quality of data. To do this, facts are randomly selected with their respective Wikipedia pages that are used as Ground Truth (GT).

DBpedia is a multilingual knowledge base built by extracting structured information from Wikipedia (*e.g.,* infoboxes) and makes this information accessible on the Web [6]. Since DBpedia is populated from Wikipedia pages in different languages, data retrieved are sometimes conflicting. To manage these conflicts, DBpedia has a module called *Sieve* which performs a quality assessment by computing some metrics such as the "recency" or the "reputation" of data before applying a fusion step based on these dimensions [13, 99].

Freebase is a graph created in 2007 which provides general human knowledge and which aims to be a public directory of world knowledge. A component included in Freebase called Mass Typer, allows a user to complete data and reconcile it semi-automatically with data already present in Freebase by performing three actions: merge, skip, or add the data. Then acquired by Google and used to support systems like Google Search, Google Maps, and Google KG, nowadays, Freebase is closed, and its knowledge has been transferred to Wikidata [9].

ConceptNet is the KG version of the Open Mind Common Sense project that contains information about words from several languages and their roles in natural language. It was built by collecting knowledge from multiple data sources namely Open Mind Common Sense, Wiktionary, games with a purpose for harvesting common knowledge, Open Multilingual WordNet, JMDict, OpenCyc, and DBPedia. Each node corresponds to a word or a sentence, and the relations between nodes are attached to numerical values that intend to represent the level of uncertainty about the relation [126].

3.3. Enterprise Knowledge Graphs

EKGs are major assets for companies since they can support various downstream applications including knowledge/vocabulary sharing and reuse, data integration, information system unification, search, or question answering [44, 109, 123]. This led companies such as Google, Microsoft, Amazon, Facebook, Orange and IBM, to build their own KGs [67, 109].

For instance, Microsoft built **Bing KG** to answer any kind of question through Bing search engine. With a size of about two billion entities and 55 billion facts according to [109], it contains general information about the world like people, or places and allows users to take actions like watching a video or buying a song. Alternatively, KGs can also increase understanding of user behavior.

It is the case of the **Facebook KG** that establishes links between the users as well as interests of users, *e.g.,* movies or music tastes. The Facebook KG is the largest social graph with about 50 million entities and 500 million statements in 2019. To handle conflicting information, the Facebook KG removes information if the associated confidence is low, otherwise, conflicting information is integrated with its provenance and the estimated confidence of the information.

Yahoo KG [106] offers different services such as a search engine, a discovery system to relate entities, or for entity recognition in queries and text. To build their KG, they leverage Wikipedia and Freebase as the backbone of the KG and use various complementary data sources to maximize the relevance, comprehensiveness, correctness, freshness, and consistency of knowledge. They mainly validate the data *w.r.t.* the ontology and through a user interface that enables entities to be corrected and updated.

[4] http://rtw.ml.cmu.edu/rtw/

Also, Orange bootstraps its KG from a set of terms of interest from a enterprise repository [67]. These terms of interest are aligned with equivalent Wikidata entities before applying an expansion to retrieve the neighborhood to extend their KG. To ensure the quality of this initial KG, pruning methods based on Euclidean distance in the embedding space, degrees of Wikidata entities, or a method based on analogical inference are used [67, 68].

In [109], the authors mention future challenges including disambiguation, knowledge extraction from unstructured and heterogeneous sources, and knowledge evolution management in the process of KG construction. We discuss some of these challenges in the next section.

4. Knowledge Acquisition

The previous section introduced some KGs including open KGs and EKGs. To build such KGs, we could rely on knowledge extraction that is the first step of knowledge integration process. In this section, we present what knowledge extraction is and some well-known automatic approaches that contribute to uncertainty in Section 4.1 that extract knowledge from: texts (Section 4.1.1), Web (Section 4.1.2), and Large Language Models (LLMs) with the recent interest in probing methods (Section 4.1.3). Finally, the definition of KG quality and related metrics are provided in Section 4.2.

4.1. Knowledge Extraction

Methods to populate a KG depend on the knowledge domain and the desired graph coverage. For example, a first method could rely on the knowledge of domain experts and populate the graph manually (*e.g.,* by crowdsourcing such as Wikidata [138] or Freebase [9]). However, this is a time-consuming process, particularly if the graph is intended to be large and it can suffer from quality issues [12, 124]. Furthermore, such open KGs have a large community that enterprises or specific KGs may not have. Therefore, large KGs such as Google, Amazon and Bing KGs rely on automatic construction methods [109]. In the following sections, we present the different tasks involved in knowledge extraction from different types of data sources.

4.1.1. From texts

For a long time, the majority of data was represented and exchanged in the form of text [95, 112]. Texts in all their forms (e.g. reports, articles, or any other textual document) are an invaluable source of information, as they are the most widely used data formats in the world (*e.g.,* the scientific research area, where knowledge is communicated via scientific articles [65]). To leverage knowledge from texts as data sources to enrich a KG, we rely on the task called Information Extraction (IE) (or knowledge acquisition). IE transforms unstructured information in text form into structured information, *i.e.,* $(entity_1, relation, entity_2)$ triples [107]. The aim of information extraction is to identify entities, their attributes, and their relationships with other entities in text [148]. In general, this task is separated into several sub-tasks: Entity Recognition (ER) or Named Entity Recognition (NER) and Relation Extraction (RE). Figure 4 depicts the input and output of a text-based knowledge extraction task. NER aims to identify named entities into the text and classify them to general types, while RE extracts semantic relationships that occur at least between two entities [49, 167].

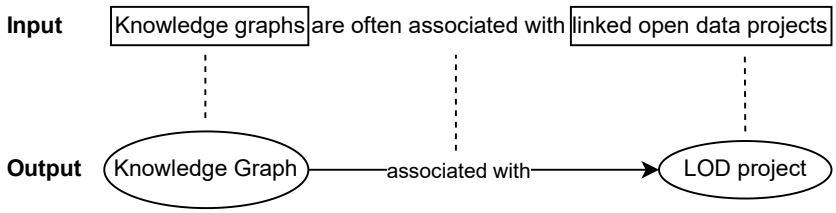

Fig. 4. Illustration of knowledge extraction from a single sentence. LOD stands for Linked Open Data.

There are two main IE approaches in the literature [107]: Traditional Information Extraction (Traditional IE) and Open Information Extraction (Open IE). Traditional IE relies on manually defined extraction patterns or patterns

learned from manually labeled training examples [107]. However, if the domain of interest evolves, the user must redefine the extraction patterns. Open IE does not rely on predefined patterns and faces three challenges [107, 157]: automation, text heterogeneity and scalability. Automation means that the information extraction system must rely on unsupervised extraction strategies. Heterogeneity stands for the difference of text types across knowledge, *e.g.,* a scholar journal versus a popular science journal. Because the extractions are performed in an unsupervised manner *i.e.,* without any labelling or predefined schema to support the extraction, it implies higher uncertainty in the extracted knowledge. Furthermore, as text types are heterogeneous due to its unstructured form, knowledge extraction patterns are more general and can lead to different levels of specificity of knowledge. Finally, the system must be able to handle large volumes of text for scalability reasons. The most widespread approaches to tackling these issues consist of pipelines composed of methods based on Natural Language Processing (NLP) [148].

One of the earliest examples of a Traditional IE systems is **KnowItAll** [37] that automates the domain-independent extraction of large collections of facts (*i.e.,* triples) from the Web. It is nevertheless supported by an extensible ontology and a minimal set of generic rules to extract entities and relations contained in its ontology. KnowItAll includes four components: an extractor, a search system, an evaluator, and a database. Its extractor instantiates a set of extraction rules for each class and relation based on a generic domain-independent pattern, for example *"cities such ..."* $\rightarrow$ *"cities such as Paris, Stockholm, ..."* deduces that Paris and Stockholm are instances of a "City" class. The search component, which includes 12 search engines such as Google, applies queries based on the extraction rules, *i.e., "Cities such as"*, then retrieves the web pages and applies the extractor. An evaluator leverages the statistics provided by search engines to assess the probability that the extracted relationships are valid. Once the extracted data has passed through these three components, it is stored into a relational database.

Traditional approaches for information extraction rely on an extractor for each target relation based on labeled training examples (*e.g.,* pre-designed extraction patterns). However, these approaches do not address the problem of extraction on large corpora whose relations are not all specified in advance [40] whereas Open IE no longer relies on predefined patterns and allows new information to be explored [107].

For example, **TextRunner** [157] that introduced the concept of Open IE, extracts a set of relational tuples without human input required. TextRunner is described by three components. The first one is a single-pass extractor that labels the text with part-of-speech tags (PoS) (*i.e.,* grammatical tagging) and extracts (e_1, r, e_2) triples. The second component is a self-supervised classifier trained to detect the correctness of the extraction. Finally, the last component is a synonym resolver that groups together synonymous entities and relations, since TextRunner has no predefined relations on which extractions are guided.

A slightly more recent approach is **ReVerb** [40]. Using constraints, this method aims to resolve the inconsistent extractions of previous Open IE models due to predicates composed of a verb and a noun. Two types of constraints are introduced on relational sentences: a syntactic constraint and a lexical constraint. Firstly, the syntactic constraint imposes the relational sentence to start either with a verb, a verb followed by a noun, or a verb followed by nouns, adjectives, or adverbs. Regarding the lexical constraint, it focuses on relations that can take many arguments and not on very specific relations. According to the results, these additional constraints allow ReVerb to outperform TextRunner. In addition, ReVerb assigns a confidence score to extractions from a sentence by applying logistic regression classification. To do this, extractions of the form (x, r, y) from a sentence s and for 1000 sentences were labeled as valid or invalid, and 19 features such as "s begins with x", "x is a proper noun", "(x, r, y) covers all words in s" were used as input variables for the logistic regression model. Such confidence scores can be used for downstream knowledge extraction tasks to support their integration into the KG (see Section 5.2).

OLLIE [96] expands the syntactic scope of relations phrases to cover much larger number of expressions and allows additional context information such as attribution and clausal modifiers. The authors argue that other models lack context on extracted relations. Hence, compared to previous methods, OLLIE introduces a new component that analyzes the context of an extraction when the extracted relation is not mentioned as factual in the text. This context is attached to each extracted relation and models the validity of the information expressed (*e.g.,* mentions of "*according to*" in a sentence). In [66], the authors present multiple components involved in different IE pipelines in the literature. They propose several combinations of these components and evaluate them in a complete pipeline that includes four steps in the PLUMBER framework: Coreference Resolution, Triples Extraction, Entity Linking and Relation Linking. 40 reusable search components are combined, representing 432 distinct information extraction pipelines. Further information is provided in [66].

4.1.2. From the Web

The Web contains an huge amount of data. It is probably the most widely used tool for exchanging knowledge between people (*e.g.,* in the form of HTML texts). Therefore, it represents an invaluable data source for building KGs. However, the latter suffers from uncertain facts, in part due to the fact that anyone can edit it. In this context, it is necessary to select reliable data sources from the Web and to implement approaches for assessing the reliability of the extracted knowledge. In this section, we present some KGs that have been built from the Web.

NELL [100] has been extracting facts from the Web continuously since January 2010, and aims to improve over time. NELL is a system that takes an initial ontology as input and reads facts from the Web and removes the incorrect ones from a set of labeled data and user feedback on the trustworthiness of the extracted facts. The core of NELL consists of learning thousands of tasks to classify extracted noun phrases into categories, to find the confident relations for each pair of noun phrases, and to identify synonymous noun phrases. Then, the extracted facts are stored in a KB with their provenance and confidence score computed during the relation classification step.

Knowledge Vault [35] is a probabilistic KB that combines extractions from Web content and prior knowledge derived from existing knowledge repositories such as Freebase. They rely on the Local Closed World Assumption, *i.e.,* for a set of existing object values $O(s,p)$ from an existing KG that contains a set of (s,p,o) triples, a candidate triple (s,p,o) is correct if $(s,p,o) \in O(s,p)$. However, if $(s,p,o) \notin O(s,p)$ and $|O(s,p)| > 0$, the triple is incorrect. Hence, this assumption can be difficult to adopt in the construction of an EKG. To merge the extractors (four different fact extraction methods: text documents, HTML trees, HTML tables, and Human Annotated pages) they define a feature vector f for each extracted triple, then apply a binary classifier to compute the probability that the fact is true. They assume that the confidence scores from each extractor are not necessarily on the same scale. Therefore, to cope with this issue, they apply a Platt scaling method that fits a logistic regression model to the confidence scores in order to obtain a probability distribution. Concerning the fusion task, they construct a feature vector f for each extracted triple and apply a binary classifier to compute the probability of the fact to be true given the feature vector. Each predicate is associated to a distinct classifier. Each feature vector contains the square root of the number of sources where the extractor extracted this triple and the mean score of the extractions from this extractor.

Probase [147] does not consider knowledge extracted from the Web to be deterministic, but models it using probabilities. The authors argue that existing KBs and taxonomy construction methods do not have sufficient concept coverage for a machine to understand the text in natural language. Probase includes the uncertainties of the extracted knowledge (specifically vagueness and inconsistencies that are due to the knowledge and to flawed construction methods). It was built from 1.6 billion web pages from an iterative learning algorithm that extracts pairs (x,y) that verify an *isA* relation between x and y, then a taxonomy construction algorithm organizes these extracted pairs into a hierarchy. In Probase, facts have probabilities that measure their plausibility and typicality. Plausibility is computed from multiple features *e.g.,* the PageRank score, the patterns used to extract *isA* pairs, or the number of sentences where x or y is present with its respective role (sub or super concept). Typicality is then computed as a function of plausibility and the number of evidences of the fact, *i.e.,* the number of sentences in which the fact is mentioned.

4.1.3. Probing

With the arrival of Deep Learning (DL) models and Large Language Models (LLMs), some triple extraction tasks are now successfully carried out by such models. [103] reviews some of them such as Graph-Based Neural Models, CNN-based model, Attention-Based Neural model and others applied to a specific knowledge domain. Also, with significant advances in LLMs and the fact that they are trained on a wide variety of information sources, some researchers have shifted their attention to KG construction by leveraging the knowledge learned by LLMs [112]. For example, a workshop on KB construction from pre-trained language models (KBC-LM[5]) and a challenge on language models for KB construction (LM-KBC[6]) are now proposed at the International Semantic Web Conference (ISWC).

In [49] the authors use the BERT model for NER and RE tasks to build a biomedical KG. In [48], the authors exploit knowledge encoded in LLM parameters (*a.k.a.* parametric knowledge [112]) to feed a KG by harvesting

[5] https://lm-kbc.github.io/workshop2024/
[6] https://lm-kbc.github.io/challenge2024/

knowledge for relations of interest. To illustrate their method, they provide an example of knowledge extraction for the "*potential_risk*" relation. The input contains a prompt such as "The potential risk of A is B" with a few shot of seed entity pairs that validate the relationship, *e.g., (eating candy, tooth decay)*. Then, the entity pairs obtained at the output of the LLM are ranked according to a consistency score computed *w.r.t.* the compatibility scores between entity pairs. From such point of view, knowledge can be directly extracted from models without visible difficulty by performing aggregation of conflicting information.

However, Pan *et al.* [112] explore possible interactions and synergies between KGs and LLMs including the construction of KGs from LLMs and raise several issues. LLMs can be used to extract knowledge directly, but this mainly applies to generic, non-specific domains and perform poorly on specific domains. They also lack accuracy in numerical facts such as the birthdate of a person and knowledge of long-tail entities, or have difficulty to memorize them. In addition, LLMs are subject to various biases (*e.g.,* gender bias) that are inherent to training data. Finally, LLMs do not provide any provenance or reliability information of extracted knowledge [112], which can be an obstacle for many knowledge fusion approaches presented in Section 8. In [170], the authors evaluate the ability of LLMs, in particular GPT-3.5, ChatGPT and GPT-4, on KG construction and reasoning (*i.e.,* link prediction and question answering) tasks under different settings, namely zero-shot and one-shot. The authors also point out LLMs fail to outperform state-of-the-art models for KG construction and have limitations to recognize long-tail knowledge.

4.2. Quality and metrics

Assessing the quality of the KG constructed is important since it is practically impossible to obtain a perfect KG, especially when this latter is very large and populated by automatic approaches from multiple data sources or by manual approaches where human contributors are not necessarily familiar with KGs and have different levels of expertise. Furthermore, the world is uncertain and knowledge is constantly evolving. To evaluate a KG, we can rely on five quality dimensions [59, 60, 143, 153]: completeness, accuracy, timeliness, availability, and redundancy.
Completeness refers to the coverage of knowledge for the specific domain the KG is intended to represent.
Accuracy corresponds to the correctness of facts in the KG. In [145], Weikum *et al.* define some metrics to assess the quality of a KB such as *precision* that captures the accuracy (these terms are sometimes used to describe the same thing), and *recall* that captures the completeness, in the following way:

$$precision(S) = \frac{S \cap GT}{S} \quad \text{and} \quad recall(S) = \frac{S \cap GT}{GT}$$

where S is a set of statements from the KB to be evaluated, and GT is the ground-truth set for the domain of interest. To deal with uncertain statements that are associated with a confidence score, a threshold is chosen, for which all statements with a score above this threshold are kept. They also provide an evaluation method that involves uniformly taking a sample of statements and representative of the KB and evaluating it, for example manually, where several annotators may be involved and a consensus or large majority must be found for each annotation.
Timeliness represents how up-to-date the KG is. The KG can contain temporal facts or facts that evolve and are valid only over a fixed period of time.
Availability measures the access to KG data, involving its querying and representation.
Redundancy assesses whether different statements express the same fact, which may entail an entity resolution task.

Another aspect of data quality is the preservation and representation of its provenance and certainty in the form of metadata, which can be used for questions of data selection in relation to both source and quality. The metadata can also support knowledge fusion approaches by taking them as prior knowledge, as we describe in Section 8. Other metrics are proposed in the survey [143] for each quality dimension.

5. Knowledge Graph refinement

As presented in Section 4, several approaches can be used to extract knowledge from heterogeneous sources (*e.g.,* tables, texts, databases, or human effort) to populate a KG. The advantage of using multiple sources is two-fold:

to ensure knowledge coverage and to identify inconsistencies by leveraging collective wisdom [33]. However, the world is uncertain and data sources are of varying quality leading to uncertain knowledge, which we need to handle in the integration process *w.r.t.* quality dimensions listed in Section 4.2. The causes of uncertainty are presented in Section 5.1. Then, we describe our theoretical data integration pipeline for dealing with knowledge uncertainty to enrich a KG in Section 5.2.

5.1. Knowledge Deltas

The uncertainty is everywhere in information and can take the form of knowledge deltas between data sources, according to [29], we adopt this definition of uncertainty in this survey. We distinguish two types of uncertainty: epistemic, *i.e.,* knowledge about a piece of information is incomplete or unknown; and ontic, *i.e.,* uncertainty is inherent in the information [139]. The possible causes of uncertainty are [1, 139]: *(i)* a lack of knowledge; *(ii)* a semantic mismatch or a lack of semantic precision and *(iii)* a lack of machine precision.

Indeed, when building a KG from heterogeneous sources, some kind of deltas of knowledge between sources may appear. These deltas of knowledge can occur between two data sources on the same subject, *e.g.,* differences in granularity and contradictions. It is also possible that a data source contradicts itself, a possible way to detect these deltas is to compare the data source to itself by "reflecting on data patterns or extrapolation to complete missing information and/or detect wrong ones" according to [28]. On the other hand, duplicates can also occur if the two data sources provide exactly the same knowledge, which needs to be managed for reasons of scalability and KG quality.

We use the same definitions of knowledge deltas as [28], and illustrate them with examples from Wikipedia and Wikidata, depicted in Figure 5. Suppose that t is a statement. Among the possible knowledge deltas, we find six causes:

- **Invalidity:** t is invalid. As illustrated in Figure 5 (a), the Wikipedia text of the figure provides incorrect information: the date of renaming of the Paris region to "île-de-France" is invalid in the Wikipedia page[7];
- **Vagueness:** t provides vague, imprecise information. As depicted in Figure 5 (a), the date mentioned on Wikipedia is more vague than the date provided by Wikidata[8] for the "located in the administrative territorial entity" property which contains additional information such as the day, month, and year;
- **Fuzziness:** t states a fuzzy truth, where the range of values is imprecise itself. As depicted in Figure 5 (b), the Wikipedia article[9] about 5G claims that the network has higher download speeds peaking at 10 Gbit/s without specifying a lower bound;
- **Timeliness:** a data source can provide the statement t which is no longer valid at the current time, unlike another source, which can provide an updated version of t. As in Figure 5 (c), on the Wikipedia page[10] of May 10, 2022, "Twitter" has not yet been renamed to "X". This information has now been changed, otherwise there would have been an updating issue;
- **Ambiguity:** t provides multiple interpretations. As shown in Figure 5 (d), Mercury[11] can be a planet, an element, or a god in mythology;
- **Incompleteness:** t provides incomplete information. As in Figure 5 (e), the track listing of the album "Evolve" by the group Imagine Dragons on Wikidata[12] contains fewer songs than in Wikipedia[13].

The creation of knowledge deltas can be involuntary or voluntary. An involuntary delta could be the result of uncertain knowledge about a domain (*e.g.,* popular science article *vs* expert article), a typing error, or a data source not up-to-date. A voluntary delta could simply stem from sabotage by a malicious person (for example, spreading fake news). Deltas are closely related to the quality dimensions of a KG, since they have a direct impact on them.

[7] https://en.wikipedia.org/w/index.php?title=Paris&oldid=1197869134
[8] https://www.wikidata.org/w/index.php?title=Q90&oldid=2058313448
[9] https://en.wikipedia.org/wiki/5G
[10] https://en.wikipedia.org/w/index.php?title=Twitter,_Inc.&oldid=1087087372
[11] https://en.wikipedia.org/wiki/Mercury
[12] https://www.wikidata.org/w/index.php?title=Q29868187&oldid=2009666363
[13] https://en.wikipedia.org/w/index.php?title=Evolve_(Imagine_Dragons_album)&oldid=1197244329

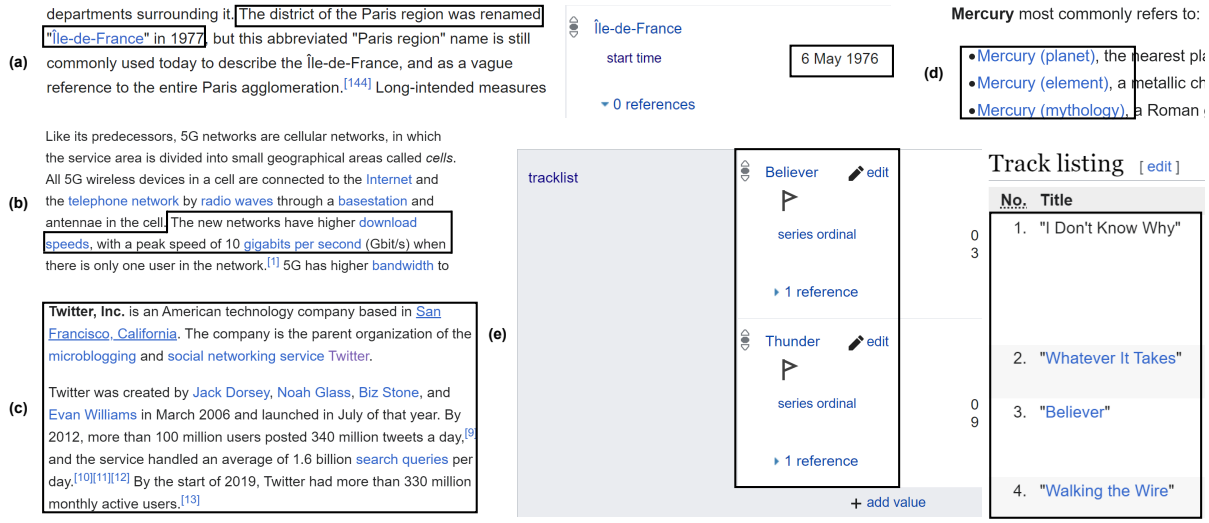

Fig. 5. Illustration of the different possible deltas about some topics between English Wikipedia and Wikidata: (a) invalidity + vagueness, (b) fuzziness, (c) timeliness, (d) ambiguity, and (e) incompleteness.

For example, a delta due to the invalidity of an information from a data source directly affects the accuracy of a KG. We propose to classify these kinds of deltas which lead to conflicts in two classes, namely *Granularity* that stands for a difference between two data sources in the specificity of knowledge and *Contradictory* that stands for an incompatibility of knowledge as depicted in Figure 6. We classify *Fuzziness*, *Incompleteness*, and *Vagueness* deltas in the granularity category. These deltas lead to different levels of specificity between knowledge of two data sources. This knowledge is not necessarily false, but may be in conflict *e.g.,* a city *vs.* a country to describe the location of an event. On the other hand, *Invalidity*, *Ambiguity*, and *Timeliness* deltas lead to contradictory knowledge, where some parts of the knowledge are necessarily false.

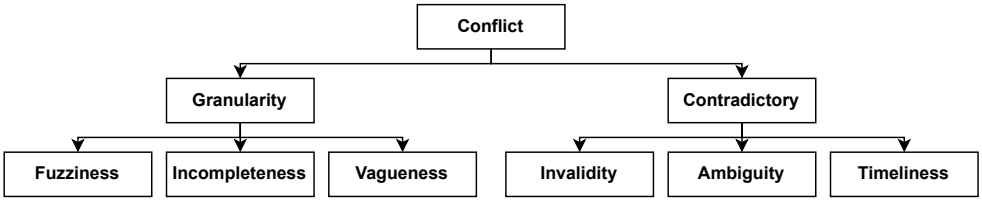

Fig. 6. Our proposed classification of knowledge deltas by type of resulting conflict (difference of granularity or contradiction).

In [4, 160], the authors assume that uncertainty is a common feature of the knowledge we handle daily. In this sense, exploiting uncertain data sources by ignoring uncertainty to enrich a KG would impact downstream applications of the graph. The life-cycle for exploiting uncertain data sources requires the measure, the quantification, and the integration of uncertainty in the KG. In such a view, the uncertainty should be considered anywhere in the pipeline of data integration, including its representation within the KG. We present our ideal data integration pipeline that tackle the aforementioned requirements in the following section.

5.2. Requirements for an ideal data integration pipeline

All ways of enriching a KG (*e.g.,* crowdsourcing, extraction from texts or tables, etc.) are error-prone methods since a human cannot be an expert on every domain involving mistakes and extraction algorithms rarely achieve perfect accuracy/precision. Errors can occur at several stages in the data integration process that encompasses extraction, alignment, or fusion. Probably one of the most natural ways of capturing and quantifying uncertainty

caused by knowledge deltas or the reliability of knowledge integration components is to use confidence scores. As mentioned in Section 4, several extraction approaches provide confidence scores about the triples they extract. For example, each triple outputted by ReVerb [40] is associated to a confidence score obtained from a logistic regression. Another work [86] focuses on estimating a confidence score for the slot filling task, which consists in filling predefined attributes for entities in a KB population case. This confidence score is intended to support the aggregation of values from different slot filling systems. The authors have shown that confidence estimation improves performance of the task and that the correctness of values and estimated confidence are strongly correlated. In [146], the authors estimate confidence scores for an entity alignment task that represents the marginal probability that a set of mentions all refer to a same entity. Therefore, there is a need to consider these confidence scores and represent them as triple metadata along with their provenance information.

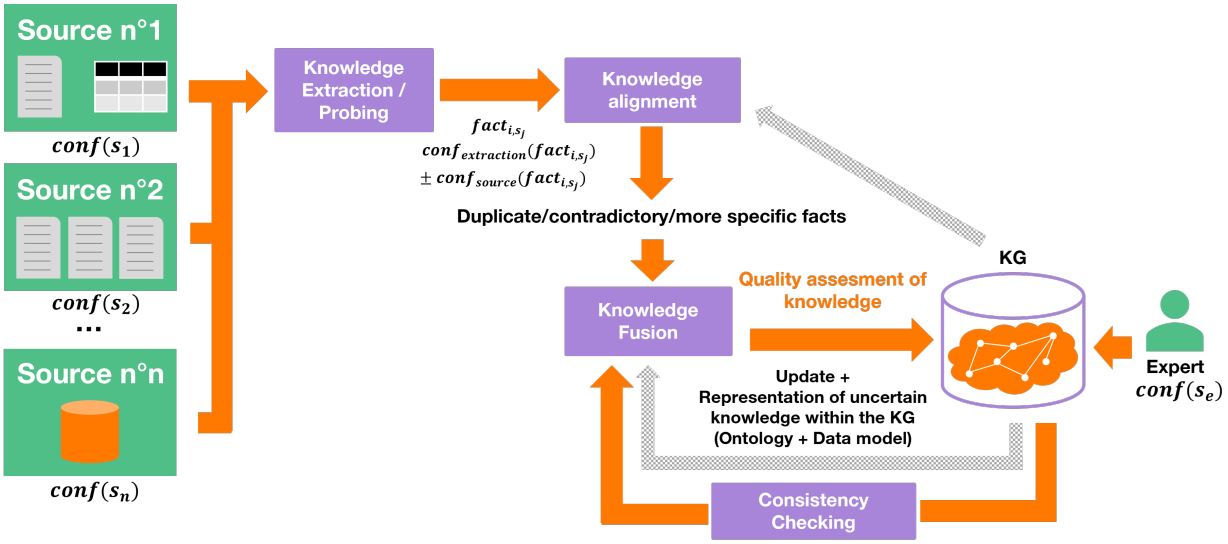

Fig. 7. Process of integrating uncertain knowledge extracted from heterogeneous sources, ideally linked to different confidence scores. Three steps form knowledge integration: (1) alignment, (2) fusion, and (3) consistency checking.

From this perspective, we propose an ideal pipeline of data integration from heterogeneous sources depicted in Figure 7. In the literature, knowledge integration after extraction is often described in two modules [8]: *Knowledge Alignment* and *Knowledge Fusion*. In this pipeline, we propose a third module called *Consistency Checking*, which actually takes place after the data integration and identifies and repairs inconsistencies in the KG, improving future knowledge enrichment. The inputs of the pipeline are multiple heterogeneous sources whose the final purpose is to feed the KG. From these data sources s_j, facts $fact_{i,s_j}$ are extracted with different confidence scores such as a confidence score in the fact by the extraction algorithm $conf_{extract}(fact_{i,s_j})$, a confidence score in the fact by the source $conf_{source}(fact_{i,s_j})$ and a confidence score in the source $conf(s_j)$. In addition to these multiple data sources, an expert can also populate the KG with a confidence score $conf(s_e)$. Before providing these facts to the KG, several tasks are required due to potential knowledge deltas. The first task *Knowledge Alignment* is the identification of duplicates, differences of granularity and contradictions among extracted facts and the KG. The next step *Knowledge Fusion* defines a policy to remove conflicting facts and keeps information as consistent, specific, and complete as possible. Then, knowledge in the KG is updated with their confidence scores and their provenance information since a user might want to query the KG about confidence of triples *w.r.t.* quality dimensions. A last step verifies the consistency within the KG. Since this step is performed after the enrichment of the KG and that this survey is focused on the uncertainty management in the construction of KGs, we do not provide further details on it in the following sections.

The aim of this pipeline is to take into account all the confidence scores in the knowledge alignment and fusion modules. This is not the case in existing work, where only the confidence scores in the data sources are leveraged

by the fusion module. However, methods for KG completion purpose (*i.e.,* predicting new relations using only the KG itself) taking into account uncertainty in an embedding space has recently been studied.

We describe and formalize the ideal data integration policy with an example. As input, we provide a set of sources $\mathcal{S}$ that contain a set of facts $\mathcal{F}$ and we obtain a $\mathcal{KG}$ as output. For each fact belonging to the source $f_{\mathcal{S}}$, this fact is aligned with the $\mathcal{KG}$. If the fact $f_{\mathcal{S}}$ conflicts with a fact already present in the KG $f_{\mathcal{KG}}$, we proceed as follows:

(1) If $f_{\mathcal{S}}$ is more specific than $f_{\mathcal{KG}}$ then we replace $f_{\mathcal{KG}}$ by $f_{\mathcal{S}}$ and we increase the confidence given to the source $\mathcal{S}$. For example, "Joe Biden is president of a North America (NA) country" "Joe Biden is President of the USA" $\rightarrow$ We keep the most specific fact, *i.e.,* "Joe Biden is president of the USA". But it would be interesting if we leverage the information "NA" by deviating from it $\rightarrow$ "USA is a country in NA" to complete the graph;
(2) Otherwise, if $f_{\mathcal{S}}$ is the duplicate of $f_{\mathcal{KG}}$, we only add the provenance of $f_{\mathcal{S}}$ and we increase the trust placed in the source $\mathcal{S}$;
(3) Otherwise, if $f_{\mathcal{S}}$ is contradictory to $f_{\mathcal{KG}}$, we resolve the conflict by finding the true value and we decrease the source that provides an erroneous fact and increase the source that provides a correct one;
(4) Else, if the fact does not conflict with a fact of $\mathcal{KG}$, the fact is added to the KG with the associated confidence scores and provenance.

A formalization of the policy is provided in Algorithm 1.

As mentioned in step (4) in the process above, we need to keep a history of knowledge provenance since the provenance information is necessary for KG quality, but could also be used for future conflict resolution or KG updating [64]. For this purpose, there is a normative ontology called PROV-O [79] that includes provenance information through three components: a set of classes, properties, and restrictions. It can be used in RDF-based KGs. The three main classes of the PROV-O ontology are *prov:Entity*, *prov:Activity*, and *prov:Agent* (*prov:Entity* is something that can be changed by an activity, *prov:Activity* is something that acts upon or with entities, and *prov:Agent* can be a human who performs an activity).

In the following sections, we detail three steps of the pipeline namely: Knowledge alignment, Knowledge Fusion, and Uncertainty Representation within the KG. We propose a section that describes uncertain KG embedding methods for KG completion and confidence prediction tasks (Section 6) before presenting the aforementioned steps. Knowledge alignment is discussed in Section 7. In Section 8, we summarize knowledge fusion methods. Finally, we explore the different mechanisms available for representing triple uncertainty in a KG in Section 9.

6. Uncertain Knowledge Graph Embedding

Embedding methods enable KG representation in a *n*-dimensional vectorial space, *i.e.,* its entities and relations are *n*-vectors. These embeddings attempt to preserve the structural properties of the graph and thus make it easy to manipulate the graph for machine learning applications such as link prediction, completion, or node classification [69]. A wide range of embedding models have emerged such as TransE [10], DistMult [154], ComplEx [137], RotatE [133], neural networks applied to graphs such as RGCN [121], or GCN [76]. In addition, embeddings are also increasingly used in the construction of KGs (for example, for knowledge alignment [42], or other tasks for KG refinement [61]). Most embedding approaches do not include the uncertainty of knowledge in their models. However, when constructing KBs, the knowledge is often uncertain or noisy and not taking into account uncertainty during the representation learning can imply a bias in its representation and impact further applications. Given the importance of embedding methods in both KG applications and construction, we consider that it is useful to gather such methods that include uncertainty expressed in terms of a confidence score in their modeling. This section describes some of these models and the datasets used to evaluate them.

6.1. Uncertain KG embedding models

In this paper, we formalize uncertain KGs as follows. An *uncertain KG* (UKG) is represented as a set of weighted triples $\mathcal{G} = (s, p, o, s_t)$, where (s, p, o) is a triple representing a fact and $s_t \in [0, 1]$ is a confidence score for this fact to be true. The uncertainty linked to the triples in the KG relies on the plausibility of the triples, but most

Algorithm 1 Data integration policy

Input: A set of facts $\mathcal{F}$ from a source $\mathcal{S}$ with confidence $conf(\mathcal{S})$ where each fact $fact_{\mathcal{S}}$ is associated to a confidence by the algorithm of extraction $conf_{extract}(fact_{\mathcal{S}})$ and a confidence by the source $conf_{source}(fact_{\mathcal{S}})$, $conf(fact_{\mathcal{S}}) = (conf(\mathcal{S}), conf_{extract}(fact_{\mathcal{S}}), conf_{source}(fact_{\mathcal{S}}))$, and a $\mathcal{KG}$

Output: $\mathcal{KG}$ updated with consistent facts of $\mathcal{S}$

```
for f_S ∈ F do
  f_KG ← align(KG, f_S, conf(fact_S))
  if f_KG ≠ ∅ then
    if moreSpecific(f_S, f_KG) then
      replace(f_KG, f_S)
      increase(conf(S))
    else if similar(f_S, f_KG) then
      addSource(KG, f_KG, S)
      increase(conf(S))
    else if contradictory(f_S, f_KG) then
      decrease(conf(S))
    end if
  else
    KG ← KG ∪ {f_S}
  end if
end for
```

KG embedding (KGE) methods do not consider this information in their modeling, making the assumption that all triples are deterministic. Such an assumption does not reflect the reality where many triples are uncertain due to the reasons described in Section 5. Table 1 summarizes the UKG embedding approaches with their associated tasks, scoring function, the year of publication, and the datasets on which the experiments were conducted. We can notice that uncertain graph embeddings have only been recently studied.

UKGE [20] improves traditional KGE models by using the Probabilistic Soft Logic (PSL) framework to infer confidence scores for unseen relational triples. Thus, UKGE encodes the KG according to confidence scores for observed and unseen triples. They map the scoring function results into confidence scores using two different mapping functions, namely a logistic function or the bounded rectifier function. For the relation fact classification, global ranking, and confidence prediction tasks, UKGE outperforms the deterministic KG embedding models such as TransE, DistMult, ComplEx and the URGE model on CN15k, NL27k and PPI5k datasets.

SUKE [141] argues that UKGE does not make full use of the structure information of the fact. Therefore, to improve this, SUKE has two components: an evaluator and a confidence generator. The evaluator defines a structure score and an uncertainty score for each fact to capture the rationality of triples. The generator outputs confidence scores for triples from the uncertainty score computed by the evaluator. The plausibility of facts are computed with DistMult [154] scoring function, then it applies a different mapping function with two parameters for the structural score and the uncertain score before merging them. The confidence generator only uses the uncertainty score to approximate the true confidence value of triples.

BEURRE [21] models entities as probabilistic boxes and relations between two entities as an affine transformation. The confidence score of the relation between two entities is represented as the volume of the intersection of their boxes. Constraints such as transitivity and composition are inserted into the modeling of embeddings to preserve these properties on relations in the embedding space. These constraints act as a loss regularization in the global loss function. Then, embeddings are trained by optimizing a loss function for a regression task and a regularization loss to apply transitivity and composition constraints.

GTransE [74] embeds uncertainty through a translational model by expanding the well known TransE [10]. The uncertainty is included at the level of the loss function on a hyperparameter of the margin loss function when training

the embeddings:

$$\mathcal{L} = \sum_{(h,r,t,s)\in Q} \sum_{(h',r,t',s)\in Q'} [f(h,r,t) - f(h',r,t') + s^{\alpha}M]_{+}$$

$$= \sum_{(h,r,t,s)\in Q} \sum_{(h',r,t',s)\in Q'} \max 0, f(h,r,t) - f(h',r,t') + s^{\alpha}M$$

where (h,r,t,s)=(head, relation, tail, confidence score), M a margin parameter and $[x]_{+}$ is the positive part of x. The scoring function f corresponds to L1 or L2-norm. Thus, with this loss function if a triple (h,r,t) has a high confidence score it will tend to respect $h+r=t$ otherwise the entity t will tend to move away from $h+r$. Before GTransE, the same authors introduced **CTransE [73]** that is closely the same model but without the hyperparameter α in power of the confidence score.

IIKE [41] models the confidence in the embedding space through a probabilistic model. The authors propose an embedding model that takes uncertainty into account by minimizing a loss function to fit the output confidence of triples acquisition (*e.g.,* NELL, or crowdsourcing) to the scoring function of the triples given by a probability function. The plausibility of a triple is modeled as a joint probability of the head of the entity, the relation and the tail $Pr(h,r,t)$ depending on $Pr(h|r,t)$, $Pr(r|h,t)$, and $Pr(t|h,r)$. For the loss function, the authors minimize the difference between the logarithm of the triple probabilities and the logarithm of the confidence of the knowledge extraction. Then apply stochastic gradient descent to refine embeddings at each iteration.

PASSLEAF [22] decomposes the model in two parts: a confidence score prediction framework that adapts the score function among existing ones, *e.g.,* ComplEx [137] or RotatE [133] and a semi-supervised learning framework.

For the UKG completion task, each relation must have enough training examples to perform correctly. **GMUC [161]** addresses the few-shot UKG completion task for long-tail relations. GMUC learns a Gaussian similarity metric that allows missing facts and their confidence scores to be predicted from a few training examples. The model encodes a support set containing a few facts with their confidence scores and a query into multidimensional Gaussian distributions. The query consists of pairs *(head, relation)* where tail and score must be predicted. Then a Gaussian matching function is used to generate a similarity distribution between the query and the support set. GMUC outperforms UKGE model on link prediction and confidence prediction on NL27K and three NL27K-derived datasets with added noise.

UOKGE [11] learns embeddings of uncertain ontology-aware KGs according to confidence scores. It encodes an instance $e_i \in E$ as a point represented by a n-dimensional vector, a class $c_i \in C$ as a sphere $s_i(c_i,\rho_i)$ where c_i is the center of the sphere and p_i is the radius, and a property $p_i \in P$ as a sphere $s_i((p_i^d,p_i^r),\varrho_i)$ where (p_i^d,p_i^r) is the center of the sphere with p_i^d representing the domain, p_i^r representing the range, and ϱ_i is the radius. Then, it introduces the following mapping function that allows changing the scale of values between 0 and 1 to represent the uncertainty. Six distinct gap functions for six types of relation are defined to encode uncertainty for: type, domain, range, subclass, sup-property and remaining properties. Then, it minimizes the mean squared error between the confidence score and those gap functions.

FocusE [111] improves KG embeddings with numerical values on edges by taking action between the scoring function of traditional models (*e.g.,* TransE, ComplEx, or DistMult) and the loss function. They introduce numerical values on edges in a way that maximizes the margin between the scores associated with true triples and their corruptions. Given the score function of an embedding model $f(t)$ with t a triple, they use a nonlinear softplus σ in order that the score provided by $f(t)$ is greater or equal to zero:

$$g(t) = \sigma(f(t)) = ln(1+e^{f(t)}) \geqslant 0$$

Then, the numerical value associated to an edge is expressed through α in the following way:

$$\alpha = \begin{cases} \beta + (1-w)(1-\beta) \text{ if } t \\ \beta + w(1-\beta) \text{ if } t^{-} \end{cases}$$

where β is a hyperparameter acting on the importance of the topological structure of the graph and w is the numerical value on the edge. Then, the final function of FocusE is the following: $h(t) = \alpha g(t)$.

ConfE [165] encodes the tuples (e, τ) with e an entity and τ an entity type by taking into account the uncertainty in each tuple. They consider entities and entity types as two different things in a KG and learn embeddings of entities and entity types in two distinct spaces with an asymmetric matrix to model their interactions. The scoring function is defined as $G(e, \tau) = e^\top M\tau$ where M is the asymmetric matrix. And uncertainty is included within the loss function:

$$\mathscr{L} = \sum_{(e,\tau)\in\mathscr{H}} \sum_{(e',\tau')\in\mathscr{H}'} \max(0, \gamma - G(e,\tau) + G(e',\tau')).C(e,\tau)$$

where $\mathscr{H}$ is the set of entities and their type and $\mathscr{H}'$ the set of corrupted tuples.

CKRL [151] introduces multiple levels of confidence, namely a local triple confidence, a global path confidence, a prior path confidence, and an adaptive path confidence. These confidence scores are integrated into the energy function following designed:

$$E(T) = \sum_{(h,r,t)\in T} E(h,r,t) \cdot C(h,r,t)$$

where $E(h,r,t) = \|h + r - t\|$ and $C(h,r,t)$ the triple confidence score aggregating all levels of confidence.

WaExt [77] embeds triples of a KG by incorporating the weight w associated to an edge (h,r,t) in the scoring function in the following manner:

$$f_w(h,r,t,w) = g(w) \cdot f(h,r,t)$$

and then minimizes a margin ranking loss function.

Wang *et al.* [142] model each entity and each relation as a multidimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ where μ is a mean vector representing its position and Σ is a diagonal covariance matrix representing its uncertainty.

MUKGE [91] aims to improve the generation of unseen facts for KGE training. The authors argue that PSL cannot take advantage of global multi-path information involving information loss to estimate the confidence of unseen facts. Indeed, PSL only considers information from simple logical rules with a path length of two as used in UKGE [20] (*e.g.,* $(college, synonym, university) \wedge$ *(university, synonym, institute)* $\rightarrow$ *(college, synonym, institute)*) and other paths in the graph between the subject and object of the inferred relation are not taken into account. To solve this issue, MUKGE introduces an algorithm called Uncertain ResourceRank to infer confidence scores for unseen triples based on the relevance of the entity pairs (subject, object). The relevance of an entity pair is computed *w.r.t.* the directed paths between subject and object in the KG. MUKGE uses circular correlation as scoring function, and either applies the sigmoid or the bounded rectifier function as the function to obtain the triple confidence. Then the authors design the loss function to fit each positive triple to its corresponding confidence score. The authors assess their model on confidence prediction, relation fact ranking, and relation fact classification. For these three tasks, MUKGE outperforms the BEURRE and UKGE models, with a focus on asymmetrical relations. However, for all relation types, performance is competitive with those of other models, except for confidence prediction, where MUKGE is the best alternative.

6.2. Datasets with numerical values on edges

In the literature, five datasets that come from uncertain KBs are commonly used in UKG completion or confidence prediction tasks presented in the previous section [111]. **CN15K** is a subset of ConceptNet (presented in Section 3.2) where the numerical values correspond to the uncertainty of the triples [126]. The confidence scores for each triple are computed *w.r.t.* the number of sources and their reliability. **NL27K** is a subset of NELL dataset (presented in Section 4.1.2) where the confidence scores are computed and refined by an Expectation Maximization (EM) algorithm and a semi-supervised learning method. **PPI5K** is a KG that represents the protein-protein interactions,

Model	Scoring function	Evaluation task	Dataset	Year
IIKE [41]	TransE	Link prediction Triple classification	NELL FB15K (synthetic)	2016
URGE[62]	Matrix Factorization (proximity preservation)	Node Clustering Node Classification k-NN search	PPI DBLP	2017
CKRL [151]	TransE	KG Noise Detection KG Completion Triple Classification	FB15K (synthetic)	2018
GTransE [74]	TransE	KG Completion	FB15K-237 (synthetic) NELL (synthetic)	2019
UKGE [20]	DistMult	Confidence prediction Relation fact ranking Relation fact classification	CN15K NL27K PPI5K	2019
UOKGE [11]	TransE	Confidence Prediction Triples Classification	CN15K	2019
SUKE [141]	DistMult	Link prediction Fact classification	CN15K NL27K PPI5K	2021
BEURRE [21]	Gumbel boxes	Confidence prediction Relation fact ranking	CN15K NELL27k	2021
PASSLEAF [22]	RotatE ComplEx DistMult	Tail Entity Prediction Confidence Prediction	PPI5K NL27K CN15K WN18RR FB15K237	2021
GMUC [161]	Minimum Similarity	Link Prediction Confidence Prediction	NL27K	2021
ConfE [165]	RESCAL	Entity Type Noise Detection Entity Type Prediction	FB15kET (synthetic) YAGO43k (synthetic)	2021
FocusE [111]	Modifiable Scoring Layer	Link Prediction (High-Valued Links)	O*NET20K CN15K NL27K PPI5K	2021
WaExt [77]	TransE TransH DistMult ComplEx	Link Prediction Triple Classification	CN15K NL27K PPI5K	2022
Wang *et al.* [142]	Similarity	Confidence Prediction Tail Entity Prediction	CN15K NL27K	2022
MUKGE [91]	Circular correlation	Confidence Prediction Relation Fact Ranking Relation Fact Classification	CN15K NL27K PPI5K	2024

Table 1

Uncertain embedding models taking into account at least one numeric value on edges with their scoring function, tasks handled, datasets on which they are evaluated, and the year of the publication.

and the numerical values correspond to the confidence relation [134]. **O*NET20K** is a dataset introduced by [111] which includes descriptions about jobs and skills. The numerical values represent the strength of the relations. Some embedding models also generate their own synthetic noisy datasets with fictitious confidence scores following probability distributions.

7. Knowledge Alignment

Knowledge alignment, *a.k.a.* knowledge resolution or knowledge matching, is the process of finding relationships or correspondences between entities of different ontologies [39]. It is the first step in the pipeline after knowledge acquisition, which identifies candidate entities for knowledge fusion. For example, in Figure 8 the entity "Galaxy S23" of both graphs refers to the same entity in real world but are stemmed from two different sources. This task copes with the "redundancy" quality dimension (Section 4.2). Whether at instance level or at ontology level, many works tackle the knowledge alignment task. This section aims to provide an overview of the knowledge alignment task and approaches by gathering the various existing surveys [38, 42, 131].

The authors of [39] distinguish different types of matching that include semantic or syntactic approaches, *e.g.,* string-based, language-based, subgraph-based, rule-based, embedding-based, or relational-based approaches. An example of such a rule-based method is the following, by [70]. For each relation R_k over the two domains, define

$$R_k(a,b) \wedge \neg R_k(a',b') \Rightarrow a \not\equiv a' \vee b \not\equiv b' \tag{1}$$

$$R_k(a,b) \wedge R_k(a',b') \Rightarrow a \equiv a' \wedge b \equiv b' \tag{2}$$

$$\forall a,b \in o,\ a',b' \in o', \tag{3}$$

where R_k is a relation present in both graphs. To reduce complexity and avoid scalability issues, some approaches use blocking methods that avoid unnecessary comparisons by gathering entities. For example, Nguyen *et al.* [104] propose different strategies of blocking based on entities' description such as *token blocking*, *i.e.,* entities in the same cluster share at least one common token in their description, *attribute clustering blocking*, *i.e.,* clusters the entities in the same group if their attributes are similar, and *prefix-infix(-suffix) blocking*, *i.e.,* exploits the pattern in the description of the URI (*e.g.,* URI infix) to create new blocks. After an optional blocking step, knowledge alignment methods are performed. Among them, [42] distinguish three methods: Sharing, Swapping, and Mapping. *Sharing* updates the entity embeddings produced by the embedding module according to the available similarity evidence of entities. *Swapping* updates the entity embeddings produced by the embedding module according to the available similarity evidence of entities but adds positive triples by leveraging aligned pairs *e.g.,* $(h,h'),\ (t,t') \rightarrow (h',r,t) + (h,r,t')$. *Mapping* learns a linear transformation between the two embedding spaces of aligned KGs.

Furthermore, alignment approaches are diverse and varied, some of them leverage attributes of entities, or use only relations between entities where different depths of context (*e.g.,* neighboring entities) are considered, while for other the path in the graph is an important aspect. We provide Table 2 that summarizes these different existing alignment approaches to get an overview strongly inspired by [42, 131].

The authors of [42] highlight that BERT_INT outperforms all models in terms of effectiveness and efficiency overall, especially when the KGs contain highly similar factual information. In fact, the alignment models that use language models such as BERT_INT are the most efficient for this task. The authors of [42] indicate the critical factors that affect the effectiveness of relation-based and attribute-based alignment methods, for instance:

- the depth of neighbors considered;
- negative sampling (for training), as the number of negatives are considered, the performances decrease;
- depending on inputs KGs to align, *e.g.,* for OpenEA datasets, is not necessary to use attributes information, factual information is sufficient.

8. Uncertain Knowledge Fusion

In the previous section, we introduced the knowledge alignment task with a summary of the embedding-based approaches that tackle it. This task identifies equivalent entities and groups them into different clusters. The next step is the fusion of the attributes of the entities into the same clusters (as illustrated in Figure 7) since they may be redundant, inconsistent, contradictory, or of different granularity. We first define the task in Section 8.1, and we present the different approaches of fusion in Section 8.2.

Model	Embedding			Method	Learning
	One-hop	Multi-hop	Path		
MTransE [18]	•			Mapping	Supervised
IPTransE [168]			•	Sharing	Semi-supervised
JAPE [129]	•			Sharing	Supervised
BootEA [130]	•			Swapping	Semi-supervised
KDCoE [19]	•			Mapping	Semi-supervised
NTAM [85]	•			Swapping	Supervised
GCNAlign [144]		•		Mapping	Supervised
AttrE [136]	•			Sharing	Supervised
IMUSE [54]	•			Sharing	Supervised
SEA [80]	•			Mapping	Supervised
RSN4EA [47]			•	Sharing	Supervised
GMNN [152]		•		Swapping	Supervised
MuGNN [15]		•		Mapping	Supervised
OTEA [115]	•			Mapping	Supervised
NAEA [169]		•		Swapping	Supervised
AVR-GCN [158]		•		Swapping	Supervised
MultiKE [162]	•			Swapping	Supervised
RDGCN [150]		•		Mapping	Supervised
KECG [81]		•		Mapping	Supervised
HGCN [149]		•		Mapping	Supervised
MMEA [125]	•			Sharing	Supervised
HMAN [155]		•		Mapping	Supervised
AKE [89]	•			Mapping	Supervised
RREA [94]		•		Sharing	Supervised
BERT_INT [135]		•		Sharing	Supervised
MTransE+RotatE [132]	•			Sharing	Supervised

Table 2

Recent embedding-based approaches that tackle knowledge alignment task.

8.1. Task definition

The knowledge fusion step consists of studying how to combine various information about the same entity or concept from multiple data sources into a consistent and a unified one regarding the different deltas listed in the Section 5.1 [59, 108]. The authors of [31] identify three broad goals to be achieved for this challenging task:

- Completeness: measures the expected amount of data (number of tuples and number of attributes) at the output of the fusion task;
- Conciseness: measures the uniqueness of object representations in the integrated data (number of unique objects and number of unique attributes of objects);
- Correctness: measures the correctness of data, *i.e.,* its conformity to the real world.

Therefore, data fusion corresponds to resolve conflicts from data with respect to these three goals. In [7], the authors distinguish two types of data conflicts from a data fusion perspective: *contradictions* and *uncertainties*. He defines *contradictions* as follows: "a contradiction is a conflict between two or more different non-null values that are all used to describe the same property of an object" and *uncertainties* as follows: "an uncertainty is a conflict

between a non-null value and one or more null values that are all used to describe the same property of an object". We adopt the same definition of *contradictions*, but we do not consider the same definition of *uncertainty*. We define the second type of conflict as a difference in granularity of knowledge, as illustrated in Figure 6.

In this survey, "uncertainty" is a more generic term whose sources lie in knowledge deltas and the inaccuracy of each step in the knowledge integration pipeline, including knowledge acquisition. Indeed, when we integrate knowledge from several sources, the quality of the information varies, and we need to determine the trustworthy information by performing a Truth Inferring (TI) task.

According to Rekatsinas [117] different TI strategies can be adopted. There are simple strategies that estimate the true values of the entities compared to the other values provided by the sources by applying a majority vote or an average on them. Then, there are strategies that use the trustworthiness of the sources to quantify the true values of the objects, and it is even possible to establish a precision for each class of object and for each source. The problem with simple strategies is that they do not take into account the varied quality of data sources [84], but these are often used to initialize true values to start iterative TI methods.

For example, we consider that we have previously extracted knowledge in triple form *(subject, predicate, object)* about the mobile phone "Galaxy S23" from several sources $\mathcal{S}_1$, $\mathcal{S}_2$, ..., $\mathcal{S}_n$ resulting in the table at the bottom of Figure 8 where entities have already been aligned. We also represent the table as a graph for sources $\mathcal{S}_1$ and $\mathcal{S}_2$. Several papers on data fusion use the term "data item", *i.e., (entity, attribute, value)* instead of the term "triple"

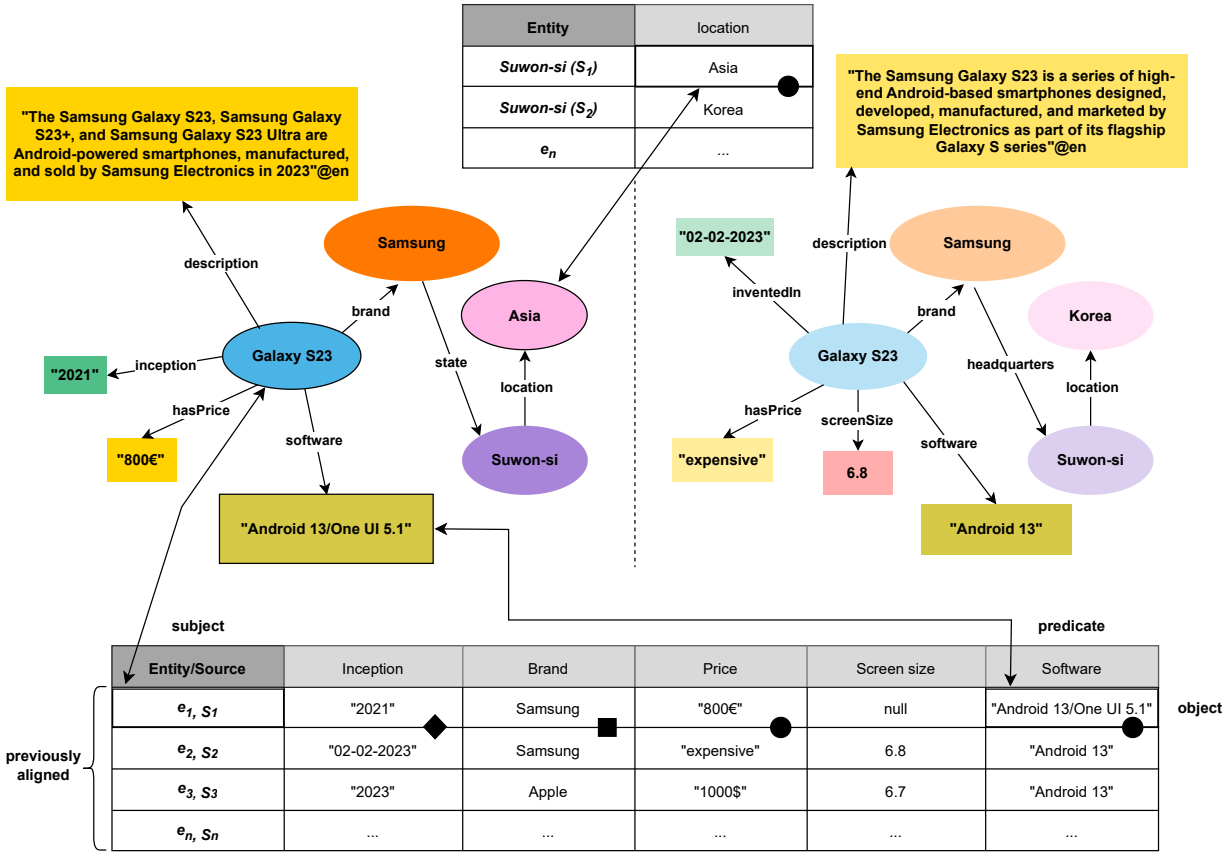

Fig. 8. Extracted triples represented as graphs from two sources $\mathcal{S}_1$ and $\mathcal{S}_2$. $\mathcal{S}_1$ on the left and $\mathcal{S}_2$ on the right. With different types of conflicts: contradictory information (◆), duplicate (■), and difference of granularity (●).

to refer to an element to be merged. However, in practice a data item is equivalent to a triple *(subject, predicate, object)*. Each row of the table corresponds to an entity of a graph and its attributes, for example the entity e_1 corresponds to the node "Galaxy S23" of the graph and the values associated with e_1 in the table are the objects of

the triples. These objects are linked to the subject "Galaxy S23" by the predicates identified by column headers, as depicted in Figure 8. The data extracted from both sources is almost the same except for the relations *software* and *inception* where differences of granularity appear (*e.g.,* the price of the mobile phone). Source $\mathcal{S}_3$ states that the brand of the mobile phone is "Apple", contradicting the first two data sources. Another example of contradiction is the invalidity of the phone's inception date provided by $\mathcal{S}_1$. We can also distinguish two different levels of granularity. The first level concerns literals (*i.e.,* numerical values, strings, etc.), for example the granularity of a product description as depicted in Figure 8. The second level concerns concepts *e.g.,* Korea *vs.* Asia to indicate the location of Suwon-si as depicted in Figure 10. On top of that, different scales of knowledge are possible as illustrated in Figure 8 among triples extracted from $\mathcal{S}_1$ we have ⟨Galaxy S23, hasPrice, "800€"⟩ and from source $\mathcal{S}_2$ we have ⟨Galaxy S23, hasPrice, "expensive"⟩, the terms "800€" and "expensive" are indeed not in the same scale, one indicates the exact price while the second one gives a qualification of the price. In all cases, the first and second sources provide complementary pieces of information for other attributes. Figure 9 shows the resulting graph after the reconciliation step that includes the knowledge alignment and the fusion step where the most complete representation of the Galaxy S23 entity is produced.

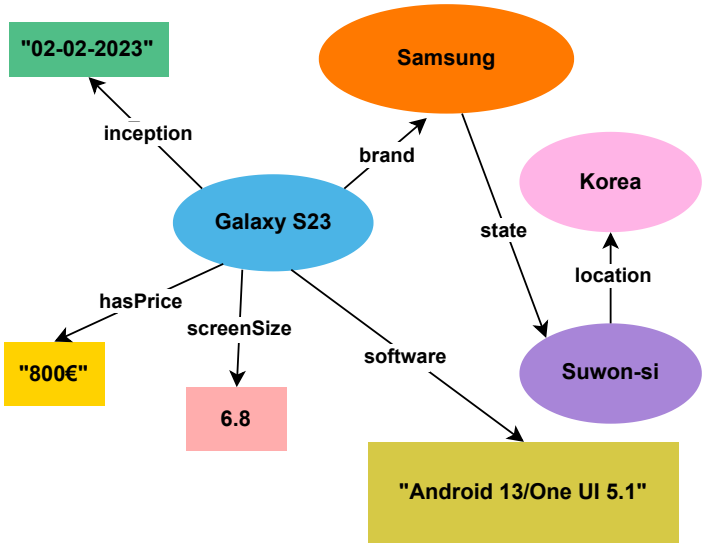

Fig. 9. Resulting KG after reconciliation step (alignment and fusion).

In the next section, we survey several methods that address knowledge fusion for truth discovery. We will use the terms "truth inferring", "truth finding" or "truth discovery" in the same way.

8.2. Fusion Approaches

Fusion methods are identified in Table 3, which provides an overview and indicates if the methods can handle some characteristics of data or data sources such as numerical data, data granularity, and dependency between data sources.

SLiMFAST [117] leverages knowledge domain features to improve the quality estimation of data sources. For example, if the data is extracted from scientific articles, the authors suggest using features such as the number of citations to the article or the year of publication, which can influence the quality of the source. Domain-specific features are included in the parameters of the logistic function that estimates source quality. To merge the data, they apply statistical learning to estimate source quality, then apply probabilistic inference to predict true values. To estimate the parameters, they use either the EM algorithm if the user does not provide labeled ground truth data, or the Empirical Risk Minimization algorithm if the user does.

ACCU [32] includes the interdependence between data sources in the truth discovery process. The intuition is that it is possible that a single source could provide the true value and that all the other sources could provide false

Model	Task		Modeling	Awareness			Datasets	Year
	TI	SQ		Numerical Data	Granularity	Source dependence		
GTM [163]	•	•	Probabilistic	•	•	○	Wikipedia edit history of city population People biographies	2012
LTM [164]	•	•	Probabilistic	○	○	○	Book author Movie director Synthetic	2012
LCA [113]	•	•	Probabilistic	•	○	○	Books Population Stocks FantasySCOTUS	2013
KBT [34]	•	•	Probabilistic	•	○	○	Triples collected by Knowledge Vault Synthetic	2015
SLiMFAST [117]	•	•	Probabilistic	•	○	○	Stocks Demonstrations Crowds Genomics	2017
LFC [116]	•	•	Probabilistic	•	○	○	Digital mammography Breast MRI	2010
DOCS[166]	•	•	Probabilistic	•	○	○	ItemCompare 4-Domain Yahoo QA SFV	2016
MDC [88]	•	•	Representation Learning	○	○	○	Dataset created from baobaozhido (crowdsourcing platform)	2017
POPACCU [33]	○	•	Probabilistic	•	○	○	Books (from AbeBooks.com) Flight	2012
ACCU [32]	•	•	Probabilistic		•	•	Synthetic	2009
CRH [84]	•	•	Loss Optimization	•	○	○	Weather Forecast (from Wunderground, HAM weather, World Weather Online) Stocks Flight	2014
Record Fusion [55]	•	○	Softmax Classifier	•	○	○	Flight Stock 1 Stock 2 Weather Address	2020
TruthFinder [159]	•	•	Probabilistic	○	•	○	Book authors	2007
ASUMS [5]	•	•	Belief functions	○	•	○	Synthetic People biographies	2016
TDH [72]	•	•	Probabilistic	•	•	○	BirthPlaces Heritages	2019
TKGC [63]	•	•	Representation Learning Probabilistic	•	•	○	Dataset built from [14] Subgraph of Freebase as prior knowledge (KG)	2022
OKELE [14]	•	•	Probabilistic	•	○	○	Synthetic	2020
FaitCrowd [92]	•	•	Probabilistic	•	○	○	SFV Dataset from crowdsourcing platform	2015
HYBRID [82]	•	•	Probabilistic	•	•	○	Book Synthetic	2017
CATD [83]	•	•	Probabilistic	•	○	○	City Population Biography Indoor Floorplan	2014
KDEm [140]	•	•	Probabilistic	•	•	○	Synthetic Population	2016

Table 3: Description of the tasks and characteristics of the data or sources that are (•) or not (○) addressed by the knowledge fusion models. (TI) stands for Truth Inference, (SQ) stands for Source Quality.

values knowing that some of them can copy on each other and therefore, spread false values. Thus, if a data source provides a value different from all the others, it is not systematically false. They define the dependency between two sources if there is a part of their data that comes directly or transitively from a common source, and it is computed by Bayesian models. Then, to discover the true value, they combine this dependency evaluation and the accuracy of the data sources which are computed in relation to the confidences of the values and dependencies between sources.

POPACCU [33], unlike ACCU, considers that the data sources are independent and that only one value can be correct.

CRH [84] (Conflict Resolution on Heterogeneous Data) estimates the reliability of a source, it uses all types of data simultaneously instead of focusing on a single one, but requires availability of all data for each entity. To initiate the source reliability score, it first applies a simple conflict resolution method, such as majority or average voting.

MDC [88] takes into account the semantic aspect of values. To illustrate the importance of semantics, in the paper the authors provide the following example: a first data source provides the true value "common cold", another source claims the value is "sinus infection" while a last source claims the value is "bone fracture", instead of examining all values at the same level, MDC calculates semantic proximity among values. This semantic proximity calculation allows evaluating how close a value is to the true value. The semantics of the values are captured by their vector representations learned by following the idea that if two values share similar words, then their vectors should be similar.

DOCS [166] is a system deployed on the Amazon Mechanical Turk that takes into account the precision of the answers of each worker to assign a specific task from a specific domain to the right worker. Regarding true value inference, the system takes advantage of the inherent relationships between the reliability of workers (which can be seen as data sources) and the true value. Thus, it considers two events: *(1)* let v be the value of an entity provided by a source s, if the quality of the values provided by s of entities in the same domain as v is high then v is likely to be correct; *(2)* then, if a source s often provides correct values for a domain d, then s has high reliability for domain d.

TruthFinder [159] claims that a fact is more likely to be true if it is provided by a reliable source, and that a source is reliable if it provides proven facts. Thus, an interdependence between facts and sources appears and consequently TruthFinder uses three elements for its iterative trust discovery process: the trustworthiness of sources, the confidence of facts, and the influences between facts. The trustworthiness $t(w)$ of a source w is computed by

$$t(w) = \frac{\sum_{f \in F(w))} s(\mathrm{t})}{|F(w)|},$$

and the confidence $s(\mathrm{t})$ of a fact t is computed by

$$s(\mathrm{t}) = 1 - \prod_{w \in W(\mathrm{t})} (1 - t(w)).$$

Then, they use the logarithm to facilitate the computation and to obtain the final scores. Although simple, these definitions integrate the influence of facts and take into account the dependence between data sources.

While most existing methods consider more generic values of a correct one as false, **TDH** [72] leverages the hierarchical structure of knowledge (*i.e.,* its one aspect of the granularity) to apply the fusion of data extracted from different sources. The idea behind is that multiple values in the hierarchy of an entity could be correct even if the predicate is functional. For example, in Figure 10, "Korea", and "Asia" are correct values for the location of "Suwon" even though one of both values is more specific. Such a modeling should not negatively impact the assessment of the reliability of the source. Instead of evaluating the value as *correct* or *incorrect*, they consider three classifications, namely *exactly correct*, *hierarchically correct*, and *incorrect*. Therefore, each source has its own generalization tendency and reliability.

In a same way, **ASUMS** [5] adapts existing truth discovery models by considering that not all values are necessarily conflicting and identifies a partial order between the values of an attribute using the "subClassOf" and "partOf" relationships. To do this, they use belief functions capable of modeling ignorance and uncertainty and allowing the incorporation of knowledge about the relations between values. Thus, in this modeling several true values can

coexist but of different granularity. Therefore, all facts more generic than a certain fact are true and conflicting facts are facts located at the same hierarchical level but with different values.

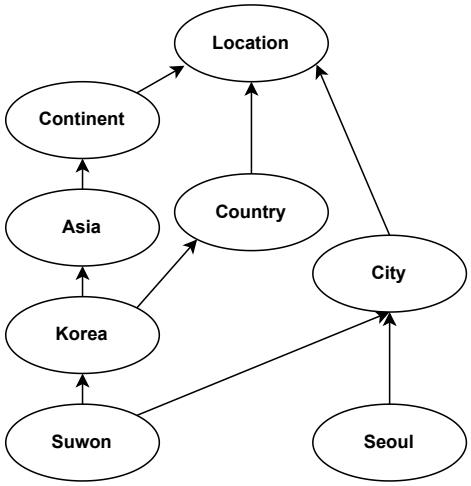

Fig. 10. Illustration of a partial ordering.

LFC [116] also measures the performance of each source (*e.g.,* annotators in the paper) by their specificity and sensitivity relating to the unknown gold standard dataset to give higher weights to the best-performing sources. The estimations are iteratively performed via the EM algorithm, where the missing data is the true value initialized by applying a majority voting. Specificity and sensitivity are then estimated in turn, and so on until convergence is reached.

LCA [113] includes four models with different sophistication, namely: SimpleLCA, GuessLCA, MistakeLCA, LieLCA. LCA is a probabilistic model where the true value is a multinomial latent variable. To infer the truth, an EM algorithm is used to compute the trustworthiness of each source with respect to the claims it makes, then the true value is computed based on the trustworthiness of the sources.

KBT [34] focuses on assessing the quality of Web sources from which facts are extracted and also evaluated. Facts are extracted as triples (subject, predicate, object) from Web pages using Knowledge Vault, which is composed of 16 different extractors. The authors extend the ACCU model by improving the estimation of the reliability of a source by distinguishing the error that comes from the fact against the error that comes from the extraction of the fact. However, they do not consider granularity or the fact that several correct values may coexist for a data item.

While other approaches are focused on categorical data, **GTM** [163] (Gaussian Truth Model) tackles the truth finding task on numerical data. This model focuses on the position of numerical claim values v_c relative to others in terms of distance to find the truth. To embed the notion of distance in their model, the authors consider the truth of each entity as a random variable and use it as the mean parameter in the probabilistic distribution for each claimed value of the observed entities. To do this, they leverage a Gaussian distribution for its ability to model errors thanks to its quadratic penalty. Regarding the evaluation of the quality of data sources, they assume that quality is related to the closeness of the claims to the truth. Therefore, the quality of a source is modeled by the variance of the Gaussian distribution, *e.g.,* a high-quality source is represented by a low variance. As aforementioned, the model can take as input the output of another truth finding method or a basic truth estimate (*e.g.,* mean or median value) to limit the effect of possible outliers on the maximum likelihood estimate (MLE). The quality of each source is determined from a prior inverse Gamma distribution, and the truth for an entity is determined from a prior Gaussian distribution. Then, to compute the truth and source quality, they perform an EM algorithm.

In the same manner as GTM, **LTM** [164] incorporates prior knowledge about sources or truth into the truth finding process and introduces the notion of two-sided source quality. It simultaneously deduces the quality of the source and the truth, as both influence each other. To compute the quality of data sources, the authors consider each source as a classifier, with its own confusion matrix. Thus, the quality of a source is defined by its sensitivity (or

recall) which corresponds to the "false negative rate" and by its specificity which corresponds to the "false positive rate" that are two independent measures. Both sensitivity and specificity are generated from a Beta distribution: the parameters for specificity are "the prior false positive count" and "the prior true negative count" and the parameters for sensitivity are "the prior true positive count" and "the prior false negative count". The prior truth probability is also modeled by a Beta distribution with the parameters "the prior true count" and "the prior false count" for each distinct (entity, value) pair. The truth value is generated by a Bernoulli distribution with a parameter θ, corresponding to the prior probability that the value is true. Finally, the truth and the quality of sources are inferred by a Collapsed Gibbs Sampling.

Record Fusion [55] merges knowledge by relying on integrity constraints, quantitative statistics, and provenance information if this latter is available. To find the true value of each table cell, they use one classifier per column (attribute) present in the table (dataset). These softmax classifiers can be modular, for example a logistic regression, a decision tree, a neural network, and so on. Three representation models are explored to create the feature vector, which will then be provided to the classifier. The first representation is the role of a cell at the column level, where three different strategies are proposed: the first acts on the format of the data (*e.g.,* the letters of the alphabet are replaced by the token "A", numbers by "N" and characters (*e.g.,* "'", ".", "'", or space) by "S" and they get a vector from a n-gram model), the second strategy is to cluster attribute values, then the last strategy consider a matrix of embeddings which will map all the values of a column (attribute) in a Euclidean space and for each cell, computes the distance between its position in space and the average position of the other values. The second concerns the role of the cell at the row level (tuple), *i.e.,* it captures the relationship of the attribute with other attributes in its row (entity). Two signals are leveraged, the first one includes the counts of pairs of attributes that are seen together, and the second one captures how often a cell occurs among rows within its own entity. Finally, the third concerns the role of the cell in relation to the complete dataset (table), *i.e.,* takes into consideration the number of denial constraint violations, includes the source information only if available since entities can have different provenance and each source can have different levels of trust. Then, the last step consists in training the different classifiers by a stage-wise additive model for a number of T iterations: *(1)* they learn the softmax classifiers with the original dataset, *(2)* use previous predictions to construct a new dynamic feature, *(3)* and learn again the classifiers using these new sets of features. For cells where the label is unknown, they assign a majority vote as their weak labels. They obtain good performance on datasets Flight, Stock, Weather, and Address about 94%-98% of precision.

In contrast to many fusion truth inferring approaches, **FaitCrowd** [92] measures the quality of data sources over several degrees, with one quality degree for each knowledge topic for a crowdsourcing case. FaitCrowd represents the expertise of each source for each topic by a Gaussian distribution. Then, it models the true value provided by a source for a certain question on a given topic as a logistic function depending on the contribution ratio of the source on the topic, the expertise of the source and a bias. To estimate the parameters, the model uses the Gibbs-EM inference method that alternates between Gibbs sampling and gradient descent.

TKGC [63] takes advantage of prior knowledge from the KG when feeding it and considers that the noise affecting the truth is represented by a probability distribution determined by the data source. To estimate the difference between the truth value and the value supplied by a source, the authors use a difference function adapted to each data type, *i.e.,* categorical data, numerical or datetime value, and string. This difference function takes as input the representation vectors previously learned in a fact scoring setting for KG completion. In fact, the probability of this latter function follows a Gaussian distribution $\mathcal{N}(0, (k_a\sigma_s^2)$ where k_a is a regularization factor and σ_s representing the noise of a data source. Then the truth inference is performed through a semi-supervised algorithm.

OKELE [14] models the probability of a fact being true by a latent random variable following a Beta(β_0, β_1) distribution, with β_1 corresponding to prior true count of the fact and β_0 corresponding to its prior false count. The quality of a data source is represented by its error variance ω_s that follows a scaled inverse chi-squared distribution Scale-inv-$\chi^2(v_s, \tau_s^2)$ representing the number of facts provided by the source v_s with variance τ_s^2. The authors argue that this distribution handles the effect of dataset size in the case of long-tail entities. The truth inference is performed by leveraging prior knowledge from existing KGs.

HYBRID [82] tackles the TI task about knowledge on tail verticals and experiment the fusion by considering two assumptions: single-truth and multi-truth. Before applying data fusion, they collect "evidences" on entities by looking for whether a source contains the subject and object of the original triple. To do this, they use three types of sources: knowledge bases (Freebase and Knowledge Vault), the Web, and query logs. The provenance information

such as the URL where the system found the evidence and the pattern are retained. Once the evidence retrieval is complete, HYBRID leverages the number of truths for each type of data items as a prior probability (for example, a mobile phone has only one year of creation, or we could consider between two and height buttons). Therefore, when a single true value or multiple true values are expected, it applies a single-truth model or a multi-truth model respectively. To assess the quality of data sources, two metrics are used: *precision*, *i.e.,* the probability when a source provides a value, a truth exists and *recall*, *i.e.,* when a truth exists, the source provides a value.

CATD [83] addresses the problem of fusion in a context where data sources provide a few claims and where estimating their quality is difficult due to the lack of data. Quality of sources are estimated by a Gaussian distribution whose mean represented the bias of the source, *i.e.,* its intentional behavior to provide false information and variance represents the reliability degree of the source. To cope with the problem of a small amount of data available from a source, they consider a confidence interval of the variance to represent their reliability. Finally, CATD applies an optimization algorithm by initializing the true values with a simple method (*e.g.,* a median of the values) and starts by estimating the quality of the sources, which depends on their claimed values, then estimates the true values.

KDEm [140] replaces the concept of true value with the concept of trustworthy opinions on the value of an entity. This model allows several true values for an entity's attribute, and consequently considers a form of knowledge granularity. KDEm leverages the kernel density estimation with a Gaussian kernel and extends it by adding the weights of the sources to estimate the probability distributions of values for each attribute of an entity. To find true values it combines the density estimation with a threshold and detects outliers that are below this threshold.

9. Uncertainty Representation

Handling knowledge uncertainty throughout the data integration process also includes its representation in the KG. Uncertainty can be represented on different value scales such as numeric, alphanumeric, textual, or intervals of values. If these different levels of uncertainty are used for reconciliation, they need to be preserved and represented in the KG as metadata in order to retain a history and could possibly be useful for resolving future conflicts [64]. The inclusion of uncertainty in the KG also enables the selection of knowledge in relation to their confidence and contributes to maintain the quality of the graph [145]. Several works deal with querying UKGs. For example, in [50], Hartig presents tSPARQL that extends RDF model and its query language SPARQL to handle uncertainty. In [24], the authors solve the failing RDF query problem, *i.e.,* when a user obtains an empty answer, that can arise when a user queries the graph with a high confidence threshold. To do this, they use tSPARQL [50] and propose answers obtained by Minimal Failing Subquery, *i.e.,* the minimal subquery contained in the failed main query, and Maximal Succeeding Subquery, *i.e.,* the maximal subquery that succeeded under the confidence threshold provided by the user. In [98], the authors propose a reasoner called URDF that solves data uncertainty for SPARQL queries. Another work tackles the task of UKG querying using UKG embeddings [43]. Therefore, it is important to choose the best knowledge representation when building a KG according to few criteria described in Section 9.2. For instance, several sources may provide the same data, hence the model must also be able to include all provenance information (*i.e.,* mutliple data sources). For specific applications, some information to assign to triples could be required *e.g.,* provenance information, the uncertainty from extraction algorithms or other, or spatial and temporal information [17, 27, 52, 105, 110]. Some formalisms of the Semantic Web offer possibilities for representing this uncertainty through metadata. Metadata is data about data defined within the RDF model and is important to estimate the validity of the information [26]. We present the uncertainty representation at the ontology level in Section 9.1 and at the data model level in Section 9.2.

9.1. Uncertainty representation at ontology level

An ontology entails three notions, namely conceptualization, explicit and formal specification, and sharing [45, 46]. The conceptualization is an abstract view of a domain, including the relevant concepts and entities, and relates them together. In this way, ontologies make domain knowledge understandable by the machine and enable reasoning about knowledge by defining rules, constraints, and the domain and range of relations [156]. [1] provides a table that summarizes the usual components that form an ontology (refer to this table for further details). OWL (Ontology

Web Language) enables to describe an ontology of a knowledge domain through individuals, classes or concepts, and properties. It is based on description logic and is part of the W3C's recommendations. Despite its ability to define rich ontologies, OWL cannot natively represent and reason about uncertainty since this latter is based on crisp logic, *i.e.,* a statement is true or false contrary to fuzzy logic for example [25]. Thus, in [23] the authors argue that the lack of ways for handling uncertain information affects the requirements of the Semantic Web. To model and handle uncertainty in an ontology by including their uncertainty theory, most approaches extend the OWL ontology [97]. These uncertainty theories include the probability theory *e.g.,* Bayesian Network (BN), Fuzzy logic, Belief functions *e.g.,* Dempster-Shafer (DS) theory.

Fuzzy-OWL [127] extends OWL with fuzzy sets theory for covering vague knowledge.

OntoBayes [156] integrates BNs into OWL in order to preserve the advantages of both. Three OWL classes are introduced: *PriorProb*, *CondProb*, and *FullProbDist* of type *ProbValue* (value between 0 and 1) to manage probabilities.

PR-OWL [23] aims to provide a probabilistic extension of OWL since the probability theory can represent uncertainty by combining Bayesian probability theory with First Order Logic. In addition to OWL, the ontology includes the statistical regularities that characterize the knowledge domain, the knowledge that is incomplete, inconclusive, ambiguous, unreliable and dissonant, then the uncertainty associated with this knowledge. It has the ability to perform probabilistic reasoning with incomplete or uncertain information conveyed through an ontology but requires RDF Reification (presented in Section 9.2) since a probabislitic model includes more than on individual (N-ary relations).

BayesOWL [25] completes OWL for representing and reasoning with uncertainty based on Bayesian Networks. The BayesOWl model includes a set of structural translation rules to convert an OWL ontology into a directed acyclic graph of a Bayesian Network. It provides the encoding of two types of probability, namely priors and pair-wise conditionals through two defined OWL classes *PriorProb* and *CondProb*. Thus, a prior probability for a concept is defined as an instance of class *PriorProb* with two properties named *hasVariable* and *hasProbValue*. Then a conditional probability is represented through an instance of class *CondProb* with the same properties as the above instance and a property *hasCondition*.

URW3-XG [78] provides an ontology as a starting point to be refined. A sentence about the world is asserted by an agent. The uncertainty of a sentence has a relation *hasUncertainty* with a *derivationType*, *uncertaintyType*, *UncertaintyModel*, and a *nature*. *UncertaintyType* includes the ambiguity, empirical uncertainty, randomness, vagueness, inconsistency and incompleteness. *UncertaintyModel* includes probability, fuzzy logic, belief functions, rough sets, and other mathematical models for reasoning under uncertainty. And *UncertaintyNature* is whether aleatoric, *i.e.,* ontic or epistemic.

Poss-OWL 2 [3] provides an extension of OWL 2 to represent incomplete and uncertain knowledge with a possibilistic viewpoint. The ontology has three main classes: *concept*, *role*, and *axiom*. *Concept* is the equivalent of concept constructor of OWL 2 with an added degree that stands for the certainty level of the concept. *Role* represents the properties of objects and data. Then, *Axiom* corresponds to the possibilistic axioms (PossTBoxAxiom and PossABoxAxiom) where each axiom is associated to a real value representing the certainty level of the axiom. The main issue of Poss-OWL 2 is that it only focus on the description of uncertainty at class-level.

Riali *et al.* [119] proposes a probabilistic extension of fuzzy ontologies in order to model vague, imprecise, probabilistic knowledge since fuzzy OWL only models vagueness. Riali *et al.* also provide a comparison of different approaches for modeling uncertainty in an ontology such as PODM [58], HyProb-Ontology [101], etc.

mUnc [29] aims to unify the different uncertainty theories within a single ontology. The ontology includes the following theories: probability, evidence of Dempster-Shafer, and possibility theory. mUnc allows publishing uncertainty theories alongside their features and computation methods. Each uncertainty theory is linked to a set of features and operators. The features correspond to the metrics on which uncertainty theory is based to indicate the degree of truth, credibility, or the likelihood of a sentence.

9.2. Uncertainty representation at data model level

The basic RDF model cannot natively inject values directly into the edges. On the other hand, there are ways and other graph representations to circumvent this limitation. To detail these graph representations, we consider that we

want to represent the uncertainty through a confidence score $s \in [0, 1]$, with 0 representing low confidence and 1 representing high confidence. In [2], the authors use 10 criteria to compare five data representation models: RDF, RDF*, Named Graph, Property Graph and their model Multilayer Graph. Among these 10 criteria, we consider that two main criteria are required to represent the confidence score and the provenance of a RDF triple. The first criterion is *edge annotation* and refers to the ability of the representation model to assign attribute-value pairs to an edge. The second one is *edge as nodes*, meaning that an edge can be referenced as multiple nodes. Therefore, we review the data representation models *w.r.t.* these two criteria and their pros and cons. We illustrate the models with the triple <JoeBiden, isPresident, UnitedStates> associated with the confidence score "0.911" in Figure 11.

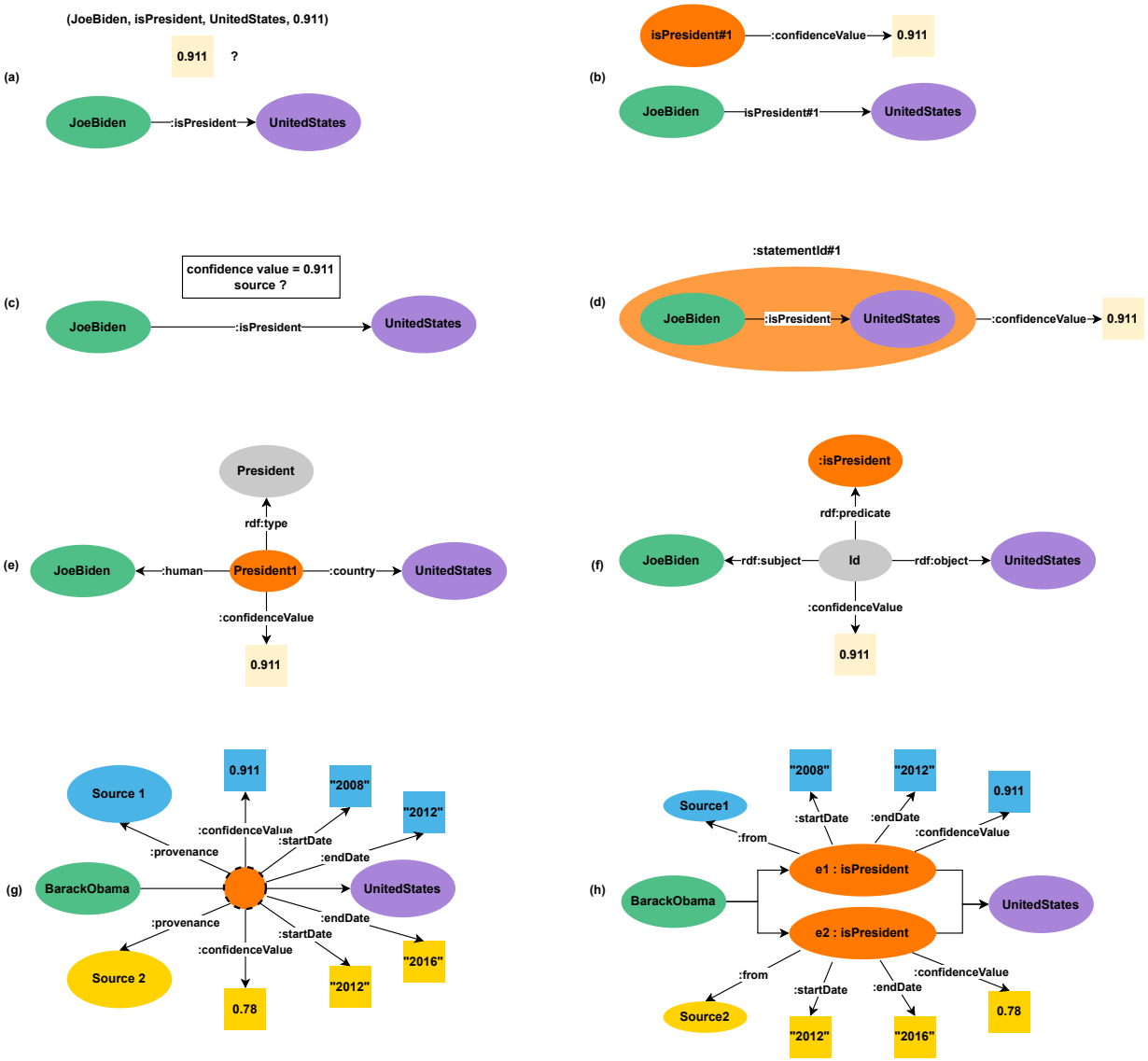

Fig. 11. Illustration of different ways to embed the uncertainty (captured by a confidence score) of a triple in the KG representation: (a) RDF, (b) Singleton Property, (c) Property Graph, (d) Named Graph, (e) N-ary, (f) RDF Reification, (g) RDF-star (RDF*), (h) Multilayer Graph. in RDF-star illustration depicts the triple <BarackObama, isPresident, UnitedStates>.

Singleton Property [105] uses a new type of property called "singleton property", which corresponds to a unique property linked to an URI between two entities. This unique property can be used as a node to which additional

relations can be added. For example, the singleton property in Figure 11(b) is "isPresident#1", then this node is used to add the confidence value "0.911". Despite the fact that the singleton property is convenient for a compact meta-level representation, this modeling introduces many unique predicates and affects data querying [52, 120].

Property Graph puts additional information about triples that are stored as a list of key/value pairs at edges in the graph. For example, in Figure 11(c), the confidence value and the provenance are attached to the relation ":isPresident".

Named Graph [56] extends the RDF triple model and allows to indicate a triple as a subgraph denoted by an IRI. This subgraph with an identifier can be used to add meta-information. In Figure 11(d), the original RDF triple <JoeBiden, :isPresident, UnitedStates> is identified by ":statementId#1", which is used as the subject in the triple <:statementId#1, :confidenceValue, 0.911>. This modeling, which corresponds to nested graphs, is well-supported in the SPARQL standard and well-suited for representing provenance data [120].

N-ary [93] creates a node to represent a relation concept whose triples linked to this node correspond to the arguments of the relation. For example, in Figure 11(e), an intermediate node "President1" of type "President" characterizes the relation ":isPresident", then meta-data can annotate the relation. The main drawback of this representation is its cumbersome syntax, which increases the complexity of the KG since the n-ary relation must be divided into several binary relations.

RDF reification [53] consists in creating an Internationalized Resource Identifier (IRI) or blank node that plays the role of the subject of all triples, as depicted in Figure 11(f). To represent the former triple it uses three new relations namely *rdf:subject*, *rdf:predicate*, and *rdf:object* then as many relations as it needs to add metadata. This modeling was the first way to make statements about statements [120]. This method is simple, but its syntax is too verbose since each statement must be reified, which considerably increases the size of the KG, makes queries and RDF data exchange more complex [2, 51, 105, 120].

RDF-star [52] is an extension of the RDF model proposed by the Semantic Web community. A RDF triple is a tuple $t^* \in (T^* \times E) \times R \times (T^* \times E \times L) \in T^*$ and RDF-star triple is a tuple $t^* \in (T^* \times E) \times R \times (T^* \times E \times L) \in T^*$ where E is the set of entities, R is the set of relations, L the set of literals, and T^* is the set of RDF-star triples. RDF-star can extend an existing RDF model expressively, since the metadata of triples are simply added as objects of them [71]. For example, in Figure 11(g), metadata such as provenance or confidence scores are added directly to the triple <BarackObama, :isPresident, UnitedStates> illustrated by  . In addition to the ability to add metadata at the statement level without modifying the remaining data, RDF-star has its own query language called SPARQL-star which reduces compatibility issues [75]. A comparison on Wikidata have shown that RDF-star performs better than reification, n-ary, and named graph representations in terms of the number of triples, loading time, and storage capacity [71]. However, RDF-star cannot consistently represent the same metadata with different values for the same triple [2, 51]. Indeed, in Figure 11(g), there are different start dates that refer to the same triple.

Multilayer Graph [2] unifies the various advantages of the other representations we have described so far into a single, simple, and flexible (whether at node or statement level) model by introducing the notion of "layer". To explain this, we use the notations and definition of a multilayer graph from [2]: given Obj a universe of objects that contains strings, numbers, IRIs and so on, a multilayer graph is defined as $G = (O, \gamma)$ where $O \subseteq Obj$ is a set of objects and $\gamma : O \rightarrow O \times O \times O$ is a partial mapping that models directed, labeled and identified edges between objects. The layers in the multilayer graph come from the nested structure of edge ids. The layer of an object $o \in O$, described as layer(o) is defined as follows: if o is not an edge id, then layer(o) = 0; otherwise if $\gamma(o) = (n_1, l, n_2)$ then layer(o) = max{layer(n_1), layer(l), layer(n_2)} + 1. Figure 12 depicts the layer representation of (h) from Figure 11(h). This data model enables to unambiguously represent multiple provenances and different confidence scores in a triple.

10. Discussion and perspectives

Throughout this survey, we have seen that several approaches exist to represent uncertainty within KGs. This is made possible by the development of ontologies that include multiple theories enabling uncertainty to be manipulated in addition to data models whose flexibility to include metadata about metadata enable additional information to be associated with extracted triples such as confidence scores. However, we can argue that methods for integrating

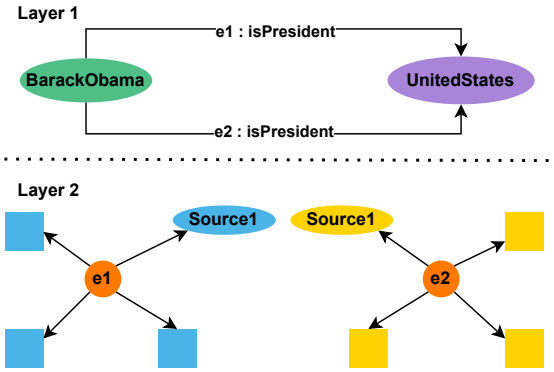

Fig. 12. Illustration of the notion of "layer" where the first layer can be seen as a single triple <BarackObama, isPresident, UnitedStates> while the second layer contains different sets of information about the triple.

knowledge after its extraction, still overlook uncertainty in their modeling, despite the recently developed methods for embedding UKGs to perform link prediction, KG completion, or confidence prediction. Taking into account the provenance information and the different levels of uncertainty operating at different locations in the knowledge integration pipeline, namely in knowledge (*i.e.,* deltas), data sources, and all components of the pipeline (*i.e.,* extraction, alignment, and fusion) would be beneficial to preserve the traceability, strengthens the quality of the KG, and enables graph querying by specifying a confidence level.

For the alignment task, the approaches do not take into account the uncertainty of the knowledge to be aligned and make the assumption that the knowledge are deterministic. On the contrary, many approaches that tackle KG completion tasks take knowledge uncertainty into account in their models. We believe that extending these models to the task of knowledge alignment would be beneficial. For example, using embedding models of uncertain graphs for mapping-based alignment methods where a transformation function between the two embedding spaces of the two KGs to be aligned is learned. Or simply include confidence scores in neighbor aggregation for GNN-based models.

Once the knowledge has been aligned, it needs to be merged. We have seen several methods dealing with different knowledge characteristics such as granularity or numerical values. Knowledge granularity is an essential aspect when building a KG from multiple heterogeneous sources. Indeed, if we leverage several popular data sources such as Wikipedia or Wikidata and one data source specific to the domain the KG is intended to represent, we are likely to face differences in granularity that we need to manage. If we use the simplest fusion approaches, such as majority voting or averaged voting, the graph will not contain the most specific knowledge. We have seen that most fusion methods do not handle this aspect of granularity, and consider that only one true value exists. Only a few methods tackle this aspect by considering a partial order between the values to be fused or a semantic distance for categorical data. We therefore recommend developing this aspect further in the modeling of fusion models, for example by estimating a granularity score for a data source in parallel with its trustworthiness score, depending on the needs of KG builders. One way of solving this problem is to further develop fusion models to capture the correlation between the attributes of the entities and to identify any inconsistencies in one or more of its attributes. Current fusion models incorporate a confidence score that embodies the trustworthiness of data sources to infer truth in knowledge fusion. Nevertheless, the confidence in extraction algorithms and other components is not accounted for. This modeling is not a problem when the same entities can be extracted from multiple sources. However, when we deal with long-tail entities and when few data sources provide knowledge about them (for example, two data sources), if one contradicts the other and their confidence score are close, the fusion model may have difficulty to find the true value while other confidence scores such as in extraction could guide the fusion model. We also advocate for better fusion models, since most proposals are specifically adapted to numerical data (which represents a relative small part of the entire data involved in KGs), only deal with categorical data, or consider both types of data but handle them in the same way. We believe that it would be more advantageous to consider the different data types in different ways, but within a single framework.

11. Conclusion

In our current world, where knowledge may be noisy, contradictory and of different granularity, uncertainty should be taken into account when constructing a KG from multiple and heterogeneous data sources. In fact, since KG construction relies on automatic knowledge extractions, other levels of uncertainty should be accounted for.

In this paper, we proposed a classification of knowledge related uncertainty into two categories: uncertainty leading to contradictions and uncertainty leading to granularity disparities. We then discussed a theoretical pipeline for the refinement of uncertain knowledge to be integrated in KG construction. This pipeline consists of four main tasks: knowledge representation (including uncertainty and provenance in the KG), knowledge alignment, knowledge fusion, and consistency checking. We also discussed challenges and perspectives on the integration of uncertain knowledge into a KG.

In particular, we have pointed out that tasks such as link prediction and KG completion are currently tackled with representational methods (embeddings) that take into account uncertainty. Knowledge alignment is a well-studied topic, with a wide range of models available from rule-based models to deep learning models, for which we provided a brief overview of existing methods. We also revisited knowledge fusion approaches, most of which based on probabilistic models, and estimated both the trustworthiness of data sources and true values. However, knowledge integration remains a challenging topic for future research. While the representation of uncertainty in a KG has received attention over the last few years (both at the ontological level and at the data model), the current knowledge integration approaches addressing both tasks remain limited in their scope (not taking into account all types of uncertainty and of knowledge deltas since they are only concerned with uncertainty).

References

[1] S.K. Anand and S. Kumar, Uncertainty Analysis in Ontology-Based Knowledge Representation, *New Gener. Comput.* **40**(1) (2022), 339–376.

[2] R. Angles, A. Hogan, O. Lassila, C. Rojas, D. Schwabe, P.A. Szekely and D. Vrgoc, Multilayer graphs: a unified data model for graph databases, in: *GRADES-NDA '22: Proceedings of the 5th ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), Philadelphia, Pennsylvania, USA, 12 June 2022*, ACM, 2022, pp. 11:1–11:6.

[3] S. Bal-Bourai and A. Mokhtari, Poss-OWL 2: Possibilistic Extension of OWL 2 for an Uncertain Geographic Ontology, in: *18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2014, Gdynia, Poland, 15-17 September 2014*, Procedia Computer Science, Vol. 35, Elsevier, 2014, pp. 407–416.

[4] D. Benslimane, Q.Z. Sheng, M. Barhamgi and H. Prade, The Uncertain Web: Concepts, Challenges, and Current Solutions, *ACM Trans. Internet Techn.* **16**(1) (2016), 1:1–1:6.

[5] V. Beretta, S. Harispe, S. Ranwez and I. Mougenot, How Can Ontologies Give You Clue for Truth-Discovery? An Exploratory Study, in: *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, WIMS 2016, Nîmes, France, June 13-15, 2016*, ACM, 2016, pp. 15:1–15:12.

[6] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann, DBpedia - A crystallization point for the Web of Data, *J. Web Semant.* **7**(3) (2009), 154–165.

[7] J. Bleiholder and F. Naumann, *Conflict Handling Strategies in an Integrated Information System*, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik, 2006. doi:http://dx.doi.org/10.18452/2460.

[8] J. Bleiholder and F. Naumann, Data Fusion, *ACM Comput. Surv.* **41**(1) (2009).

[9] K.D. Bollacker, C. Evans, P.K. Paritosh, T. Sturge and J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, ACM, 2008, pp. 1247–1250.

[10] A. Bordes, N. Usunier, A. García-Durán, J. Weston and O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, in: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2013, pp. 2787–2795.

[11] K. Boutouhami, J. Zhang, G. Qi and H. Gao, Uncertain Ontology-Aware Knowledge Graph Embeddings, in: *Semantic Technology - 9th Joint International Conference, JIST 2019, Hangzhou, China, November 25-27, 2019, Revised Selected Papers*, Communications in Computer and Information Science, Vol. 1157, Springer, 2019, pp. 129–136.

[12] F. Brasileiro, J.P.A. Almeida, V.A. de Carvalho and G. Guizzardi, Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata, in: *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, ACM, 2016, pp. 975–980.

[13] V. Bryl and C. Bizer, Learning conflict resolution strategies for cross-language Wikipedia data fusion, in: *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, ACM, 2014, pp. 1129–1134.

[14] E. Cao, D. Wang, J. Huang and W. Hu, Open Knowledge Enrichment for Long-tail Entities, in: *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, ACM / IW3C2, 2020, pp. 384–394.

[15] Y. Cao, Z. Liu, C. Li, Z. Liu, J. Li and T. Chua, Multi-Channel Graph Neural Network for Entity Alignment, in: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, A. Korhonen, D.R. Traum and L. Màrquez, eds, Association for Computational Linguistics, 2019, pp. 1452–1461.

[16] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R.H. Jr. and T.M. Mitchell, Toward an Architecture for Never-Ending Language Learning, in: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, M. Fox and D. Poole, eds, AAAI Press, 2010, pp. 1306–1313.

[17] J.J. Carroll, C. Bizer, P.J. Hayes and P. Stickler, Named graphs, *J. Web Semant.* **3**(4) (2005), 247–267.

[18] M. Chen, Y. Tian, M. Yang and C. Zaniolo, Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment, in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, C. Sierra, ed., ijcai.org, 2017, pp. 1511–1517.

[19] M. Chen, Y. Tian, K. Chang, S. Skiena and C. Zaniolo, Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment, in: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, ed., ijcai.org, 2018, pp. 3998–4004.

[20] X. Chen, M. Chen, W. Shi, Y. Sun and C. Zaniolo, Embedding Uncertain Knowledge Graphs, in: *Proceedings of the AAAI conference on artificial intelligence*, AAAI Press, 2019, pp. 3363–3370.

[21] X. Chen, M. Boratko, M. Chen, S.S. Dasgupta, X.L. Li and A. McCallum, Probabilistic Box Embeddings for Uncertain Knowledge Graph Reasoning, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Association for Computational Linguistics, 2021, pp. 882–893.

[22] Z. Chen, M. Yeh and T. Kuo, PASSLEAF: A Pool-bAsed Semi-Supervised LEArning Framework for Uncertain Knowledge Graph Embedding, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press, 2021, pp. 4019–4026.

[23] P.C.G. da Costa, K.B. Laskey and K.J. Laskey, PR-OWL: A Bayesian Ontology Language for the Semantic Web, in: *Uncertainty Reasoning for the Semantic Web I, ISWC International Workshops, URSW 2005-2007, Revised Selected and Invited Papers*, Lecture Notes in Computer Science, Vol. 5327, Springer, 2008, pp. 88–107.

[24] I. Dellal, S. Jean, A. Hadjali, B. Chardin and M. Baron, Query answering over uncertain RDF knowledge bases: explain and obviate unsuccessful query results, *Knowledge and Information Systems* **61**(3) (2019), 1633–1665.

[25] Z. Ding, Y. Peng and R. Pan, BayesOWL: Uncertainty modeling in semantic web ontologies, *Soft computing in ontologies and semantic web* (2006), 3–29.

[26] R.Q. Dividino, S. Schenk, S. Sizov and S. Staab, Provenance, Trust, Explanations - and all that other Meta Knowledge, *Künstliche Intell.* **23**(2) (2009), 24–30.

[27] R.Q. Dividino, S. Sizov, S. Staab and B. Schueler, Querying for provenance, trust, uncertainty and other meta knowledge in RDF, *J. Web Semant.* **7**(3) (2009), 204–219.

[28] A.E.A. Djebri, Uncertainty Management for Linked Data Reliability on the Semantic Web. (Gestion de l'Incertitude pour la fiabilité des Données Liées dans le Web Sémantique), PhD thesis, Côte D'Azur University, France, 2022. https://tel.archives-ouvertes.fr/tel-03679118.

[29] A.E.A. Djebri, A.G.B. Tettamanzi and F. Gandon, Publishing Uncertainty on the Semantic Web: Blurring the LOD Bubbles, in: *Graph-Based Representation and Reasoning - 24th International Conference on Conceptual Structures, ICCS 2019, Marburg, Germany, July 1-4, 2019, Proceedings*, Lecture Notes in Computer Science, Vol. 11530, Springer, 2019, pp. 42–56.

[30] X.L. Dong, Generations of Knowledge Graphs: The Crazy Ideas and the Business Impact, *Proc. VLDB Endow.* **16**(12) (2023), 4130–4137.

[31] X.L. Dong and F. Naumann, Data fusion: resolving data conflicts for integration, *Proceedings of the VLDB Endowment* **2**(2) (2009), 1654–1655.

[32] X.L. Dong, L. Berti-Équille and D. Srivastava, Integrating Conflicting Data: The Role of Source Dependence, *Proc. VLDB Endow.* **2**(1) (2009), 550–561.

[33] X.L. Dong, B. Saha and D. Srivastava, Less is More: Selecting Sources Wisely for Integration, *Proc. VLDB Endow.* **6**(2) (2012), 37–48.

[34] X.L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun and W. Zhang, Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources, *Proc. VLDB Endow.* **8**(9) (2015), 938–949.

[35] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun and W. Zhang, Knowledge vault: a web-scale approach to probabilistic knowledge fusion, in: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, ACM, 2014, pp. 601–610.

[36] L. Ehrlinger and W. Wöß, Towards a Definition of Knowledge Graphs, in: *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*, CEUR Workshop Proceedings, Vol. 1695, CEUR-WS.org, 2016.

[37] O. Etzioni, M.J. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D.S. Weld and A. Yates, Web-scale information extraction in knowitall: (preliminary results), in: *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, ACM, 2004, pp. 100–110.

[38] J. Euzenat and P. Shvaiko, *Ontology matching*, Springer, 2007.

[39] J. Euzenat and P. Shvaiko, *Ontology Matching, Second Edition*, Springer, 2013.

[40] A. Fader, S. Soderland and O. Etzioni, Identifying Relations for Open Information Extraction, in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, ACL, 2011, pp. 1535–1545.

[41] M. Fan, Q. Zhou and T.F. Zheng, Learning Embedding Representations for Knowledge Inference on Imperfect and Incomplete Repositories, in: *2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016, Omaha, NE, USA, October 13-16, 2016*, IEEE Computer Society, 2016, pp. 42–48.

[42] N. Fanourakis, V. Efthymiou, D. Kotzinos and V. Christophides, Knowledge graph embedding methods for entity alignment: experimental review, *Data Min. Knowl. Discov.* **37**(5) (2023), 2070–2137.

[43] W. Fei, Z. Wang, H. Yin, Y. Duan, H. Tong and Y. Song, Soft Reasoning on Uncertain Knowledge Graphs, *arXiv preprint arXiv:2403.01508* (2024).

[44] M. Galkin, S. Auer and S. Scerri, Enterprise Knowledge Graphs: A Backbone of Linked Enterprise Data, in: *2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016, Omaha, NE, USA, October 13-16, 2016*, IEEE Computer Society, 2016, pp. 497–502.

[45] T.R. Gruber, A translation approach to portable ontology specifications, *Knowledge acquisition* **5**(2) (1993), 199–220.

[46] N. Guarino, D. Oberle and S. Staab, What Is an Ontology?, in: *Handbook on Ontologies*, International Handbooks on Information Systems, Springer, 2009, pp. 1–17.

[47] L. Guo, Z. Sun and W. Hu, Learning to exploit long-term relational dependencies in knowledge graphs, in: *International conference on machine learning*, PMLR, 2019, pp. 2505–2514.

[48] S. Hao, B. Tan, K. Tang, B. Ni, X. Shao, H. Zhang, E. Xing and Z. Hu, BertNet: Harvesting knowledge graphs with arbitrary relations from pretrained language models, in: *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 5000–5015.

[49] A. Harnoune, M. Rhanoui, M. Mikram, S. Yousfi, Z. Elkaimbillah and B. El Asri, BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis, *Computer Methods and Programs in Biomedicine Update* **1** (2021), 100042.

[50] O. Hartig, Querying Trust in RDF Data with tSPARQL, in: *The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009, Proceedings*, Lecture Notes in Computer Science, Vol. 5554, Springer, 2009, pp. 5–20.

[51] O. Hartig, Reconciliation of RDF* and Property Graphs, *CoRR* **abs/1409.3288** (2014). http://arxiv.org/abs/1409.3288.

[52] O. Hartig, Foundations of RDF⋆ and SPARQL⋆ (An Alternative Approach to Statement-Level Metadata in RDF), in: *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017*, CEUR Workshop Proceedings, Vol. 1912, CEUR-WS.org, 2017.

[53] P.J. Hayes and P.F. Patel-Schneider, RDF 1.1 semantics. W3C recommendation, *World Wide Web Consortium* **2** (2014).

[54] F. He, Z. Li, Y. Qiang, A. Liu, G. Liu, P. Zhao, L. Zhao, M. Zhang and Z. Chen, Unsupervised entity alignment using attribute triples and relation triples, in: *Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22–25, 2019, Proceedings, Part I 24*, Springer, 2019, pp. 367–382.

[55] A. Heidari, G. Michalopoulos, S. Kushagra, I.F. Ilyas and T. Rekatsinas, Record fusion: A learning approach, *CoRR* **abs/2006.10208** (2020).

[56] D. Hernández, A. Hogan and M. Krötzsch, Reifying RDF: What Works Well With Wikidata?, in: *Proceedings of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems co-located with 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 11, 2015*, CEUR Workshop Proceedings, Vol. 1457, CEUR-WS.org, 2015, pp. 32–47.

[57] P. Heyvaert, B.D. Meester, A. Dimou and R. Verborgh, Rule-driven inconsistency resolution for knowledge graph generation rules, *Semantic Web* **10**(6) (2019), 1071–1086.

[58] E. Hlel, S. Jamoussi, M. Turki and A.B. Hamadou, Probabilistic Ontology Definition Meta-Model - Extension of OWL2 Meta-Model for Defining Probabilistic Ontologies, in: *Intelligent Decision Technologies 2016 - Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016) - Part I, Puerto de la Cruz, Spain, 15-17 June, 2016*, Smart Innovation, Systems and Technologies, Vol. 56, Springer, 2016, pp. 243–254. doi:10.1007/978-3-319-39630-9_20.

[59] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke and E. Rahm, Construction of Knowledge Graphs: Current State and Challenges, *Available at SSRN 4605059*.

[60] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, S. Kirrane, J.E. Labra Gayo, R. Navigli, S. Neumaier, A.-C. Ngonga Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J.F. Sequeda, S. Staab and A. Zimmermann, *Knowledge Graphs*, Synthesis Lectures on Data, Semantics, and Knowledge Vol. 22, Springer, 2021. ISBN 9783031007903. doi:10.2200/S01125ED1V01Y202109DSK022. https://kgbook.org/.

[61] 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 and A. Zimmermann, Knowledge Graphs, *ACM Comput. Surv.* **54**(4) (2022), 71:1–71:37.

[62] J. Hu, R. Cheng, Z. Huang, Y. Fang and S. Luo, On Embedding Uncertain Graphs, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, ACM, 2017, pp. 157–166.

[63] J. Huang, Y. Zhao, W. Hu, Z. Ning, Q. Chen, X. Qiu, C. Huo and W. Ren, Trustworthy Knowledge Graph Completion Based on Multi-sourced Noisy Data, in: *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, ACM, 2022, pp. 956–965.

[64] I.F. Ilyas, T. Rekatsinas, V. Konda, J. Pound, X. Qi and M.A. Soliman, Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale, in: *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, ACM, 2022, pp. 2259–2272.

[65] M.Y. Jaradeh, A. Oelen, K.E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker and S. Auer, Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge, in: *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*, ACM, 2019, pp. 243–246.

[66] M.Y. Jaradeh, K. Singh, M. Stocker, A. Both and S. Auer, Information extraction pipelines for knowledge graphs, *Knowl. Inf. Syst.* **65**(5) (2023), 1989–2016.

[67] L. Jarnac and P. Monnin, Wikidata to Bootstrap an Enterprise Knowledge Graph: How to Stay on Topic?, in: *Proceedings of the 3rd Wikidata Workshop 2022 co-located with the 21st International Semantic Web Conference (ISWC2022), Virtual Event, Hanghzou, China, October 2022*, CEUR Workshop Proceedings, Vol. 3262, CEUR-WS.org, 2022.

[68] L. Jarnac, M. Couceiro and P. Monnin, Relevant Entity Selection: Knowledge Graph Bootstrapping via Zero-Shot Analogical Pruning, in: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, ACM, 2023, pp. 934–944.

[69] S. Ji, S. Pan, E. Cambria, P. Marttinen and S.Y. Philip, A survey on knowledge graphs: Representation, acquisition, and applications, *IEEE transactions on neural networks and learning systems* **33**(2) (2021), 494–514.

[70] S. Jiang, D. Lowd, S. Kafle and D. Dou, Ontology Matching with Knowledge Rules, *Trans. Large Scale Data Knowl. Centered Syst.* **28** (2016), 75–95.

[71] R.T.K. Jr., A. Lehnert and G. Loh, Use Case: Ontologies and RDF-Star for Knowledge Management, in: *The Semantic Web: ESWC 2021 Satellite Events - Virtual Event, June 6-10, 2021, Revised Selected Papers*, Lecture Notes in Computer Science, Vol. 12739, Springer, 2021, pp. 254–260.

[72] W. Jung, Y. Kim and K. Shim, Crowdsourced Truth Discovery in the Presence of Hierarchies for Knowledge Fusion, in: *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, OpenProceedings.org, 2019, pp. 205–216.

[73] N. Kertkeidkachorn, X. Liu and R. Ichise, CTransE: Confidence-Based Translation Model for Uncertain Knowledge Graph Embedding, in: *Proceedings of the Annual Conference of JSAI 33rd (2019)*, The Japanese Society for Artificial Intelligence, 2019, pp. 1K4E105–1K4E105.

[74] N. Kertkeidkachorn, X. Liu and R. Ichise, GTransE: Generalizing Translation-Based Model on Uncertain Knowledge Graph Embedding, in: *Advances in Artificial Intelligence - Selected Papers from the Annual Conference of Japanese Society of Artificial Intelligence (JSAI 2019), Niigata, Japan, 4-7 June 2019*, Advances in Intelligent Systems and Computing, Vol. 1128, Springer, 2019, pp. 170–178.

[75] R. Keskisärkkä, E. Blomqvist, L. Lind and O. Hartig, Capturing and Querying Uncertainty in RDF Stream Processing, in: *Knowledge Engineering and Knowledge Management - 22nd International Conference, EKAW 2020, Bolzano, Italy, September 16-20, 2020, Proceedings*, Lecture Notes in Computer Science, Vol. 12387, Springer, 2020, pp. 37–53.

[76] T.N. Kipf and M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, in: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017.

[77] W.K. Kong, X. Liu, T. Racharak, G. Sun, Q. Ma and L.-M. Nguyen, Weight-aware Tasks for Evaluating Knowledge Graph Embeddings.

[78] K.J. Laskey and K.B. Laskey, Uncertainty Reasoning for the World Wide Web: Report on the URW3-XG Incubator Group, in: *Proceedings of the Fourth International Workshop on Uncertainty Reasoning for the Semantic Web, Karlsruhe, Germany, October 26, 2008*, CEUR Workshop Proceedings, Vol. 423, CEUR-WS.org, 2008.

[79] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik and J. Zhao, Prov-o: The prov ontology, *W3C recommendation* **30** (2013).

[80] C. Li, Y. Cao, L. Hou, J. Shi, J. Li and T. Chua, Semi-supervised Entity Alignment via Joint Knowledge Embedding Model and Cross-graph Model, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, K. Inui, J. Jiang, V. Ng and X. Wan, eds, Association for Computational Linguistics, 2019, pp. 2723–2732.

[81] C. Li, Y. Cao, L. Hou, J. Shi, J. Li and T.-S. Chua, Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model, Association for Computational Linguistics, 2019.

[82] F. Li, X.L. Dong, A. Langen and Y. Li, Knowledge verification for long-tail verticals, *Proceedings of the VLDB Endowment* **10**(11) (2017), 1370–1381.

[83] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan and J. Han, A Confidence-Aware Approach for Truth Discovery on Long-Tail Data, *Proc. VLDB Endow.* **8**(4) (2014), 425–436.

[84] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan and J. Han, Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation, in: *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, ACM, 2014, pp. 1187–1198.

[85] S. Li, X. Li, R. Ye, M. Wang, H. Su and Y. Ou, Non-translational Alignment for Multi-relational Networks., in: *IJCAI*, 2018, pp. 4180–4186.

[86] X. Li and R. Grishman, Confidence estimation for knowledge base population, in: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, 2013, pp. 396–401.

[87] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan and J. Han, A Survey on Truth Discovery, *SIGKDD Explor.* **17**(2) (2015), 1–16.

[88] Y. Li, N. Du, C. Liu, Y. Xie, W. Fan, Q. Li, J. Gao and H. Sun, Reliable Medical Diagnosis from Crowdsourcing: Discover Trustworthy Answers from Non-Experts, in: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*, ACM, 2017, pp. 253–261.

[89] X. Lin, H. Yang, J. Wu, C. Zhou and B. Wang, Guiding cross-lingual entity alignment via adversarial knowledge embedding, in: *2019 IEEE International conference on data mining (ICDM)*, IEEE, 2019, pp. 429–438.

[90] J. Liu, Y. Chabot, R. Troncy, V. Huynh, T. Labbé and P. Monnin, From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods, *J. Web Semant.* **76** (2023), 100761.

[91] Q. Liu, Q. Zhang, F. Zhao and G. Wang, Uncertain knowledge graph embedding: an effective method combining multi-relation and multi-path, *Frontiers Comput. Sci.* **18**(3) (2024), 183311.

[92] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji and J. Han, FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation, in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, ACM, 2015, pp. 745–754.

[93] F. Manola, E. Miller and B. McBride, Resource description framework (RDF) primer, *W3C Recommendation* **10**(5) (2004).

[94] X. Mao, W. Wang, H. Xu, Y. Wu and M. Lan, Relational reflection entity alignment, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1095–1104.

[95] J. Martínez-Rodríguez, I. López-Arévalo and A.B. Ríos-Alvarado, OpenIE-based approach for Knowledge Graph construction from text, *Expert Syst. Appl.* **113** (2018), 339–355.

[96] Mausam, M. Schmitz, S. Soderland, R. Bart and O. Etzioni, Open Language Learning for Information Extraction, in: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, ACL, 2012, pp. 523–534.

[97] D.L. McGuinness, F. Van Harmelen et al., OWL web ontology language overview, *W3C recommendation* **10**(10) (2004), 2004.

[98] T. Meiser, M. Dylla and M. Theobald, Interactive reasoning in uncertain RDF knowledge bases, in: *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, ACM, 2011, pp. 2557–2560.

[99] P.N. Mendes, H. Mühleisen and C. Bizer, Sieve: linked data quality assessment and fusion, in: *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*, ACM, 2012, pp. 116–123.

[100] T.M. Mitchell, W.W. Cohen, E.R.H. Jr., P.P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B.D. Mishra, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E.A. Platanios, A. Ritter, M. Samadi, B. Settles, R.C. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves and J. Welling, Never-ending learning, *Commun. ACM* **61**(5) (2018), 103–115.

[101] A.-W. Mohammed, Y. Xu and M. Liu, Knowledge-oriented semantics modelling towards uncertainty reasoning, *SpringerPlus* **5** (2016), 1–27.

[102] F.J. Navarrete and A. Vallecillo, Introducing Subjective Knowledge Graphs, in: *25th IEEE International Enterprise Distributed Object Computing Conference, EDOC 2021, Gold Coast, Australia, October 25-29, 2021*, IEEE, 2021, pp. 61–70.

[103] T. Nayak, N. Majumder, P. Goyal and S. Poria, Deep Neural Approaches to Relation Triplets Extraction: a Comprehensive Survey, *Cogn. Comput.* **13**(5) (2021), 1215–1232.

[104] H.L. Nguyen, D. Vu and J.J. Jung, Knowledge graph fusion for smart systems: A Survey, *Inf. Fusion* **61** (2020), 56–70.

[105] V. Nguyen, O. Bodenreider and A.P. Sheth, Don't like RDF reification?: making statements about statements using singleton property, in: *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, ACM, 2014, pp. 759–770.

[106] C. Ni, K.S. Liu and N. Torzec, Layered Graph Embedding for Entity Recommendation using Wikipedia in the Yahoo! Knowledge Graph, in: *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, A.E.F. Seghrouchni, G. Sukthankar, T. Liu and M. van Steen, eds, ACM / IW3C2, 2020, pp. 811–818.

[107] C. Niklaus, M. Cetto, A. Freitas and S. Handschuh, A Survey on Open Information Extraction, in: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Association for Computational Linguistics, 2018, pp. 3866–3878.

[108] A. Nikolov, V.S. Uren and E. Motta, KnoFuss: a comprehensive architecture for knowledge fusion, in: *Proceedings of the 4th International Conference on Knowledge Capture (K-CAP 2007), October 28-31, 2007, Whistler, BC, Canada*, ACM, 2007, pp. 185–186.

[109] N.F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson and J. Taylor, Industry-scale Knowledge Graphs: Lessons and Challenges, *ACM Queue* **17**(2) (2019), 20.

[110] F. Orlandi, D. Graux and D. O'Sullivan, Benchmarking RDF Metadata Representations: Reification, Singleton Property and RDF, in: *15th IEEE International Conference on Semantic Computing, ICSC 2021, Laguna Hills, CA, USA, January 27-29, 2021*, IEEE, 2021, pp. 233–240. doi:10.1109/ICSC50631.2021.00049.

[111] S. Pai and L. Costabello, Learning Embeddings from Knowledge Graphs With Numeric Edge Attributes, in: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, ijcai.org, 2021, pp. 2869–2875. doi:10.24963/ijcai.2021/395.

[112] J.Z. Pan, S. Razniewski, J. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni and D. Graux, Large Language Models and Knowledge Graphs: Opportunities and Challenges, *TGDK* **1**(1) (2023), 2:1–2:38.

[113] J. Pasternack and D. Roth, Latent credibility analysis, in: *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 1009–1020.

[114] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, *Semantic Web* **8**(3) (2017), 489–508.

[115] S. Pei, L. Yu and X. Zhang, Improving cross-lingual entity alignment via optimal transport, International Joint Conferences on Artificial Intelligence, 2019.

[116] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni and L. Moy, Learning From Crowds, *J. Mach. Learn. Res.* **11** (2010), 1297–1322.

[117] T. Rekatsinas, M. Joglekar, H. Garcia-Molina, A.G. Parameswaran and C. Ré, SLiMFast: Guaranteed Results for Data Fusion and Source Reliability, in: *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, ACM, 2017, pp. 1399–1414.

[118] D. Reynolds, Position paper: Uncertainty reasoning for linked data, in: *Workshop*, Vol. 14, 2014.

[119] I. Riali, M. Fareh and H. Bouarfa, A Semantic Approach for Handling Probabilistic Knowledge of Fuzzy Ontologies, in: *Proceedings of the 21st International Conference on Enterprise Information Systems, ICEIS 2019, Heraklion, Crete, Greece, May 3-5, 2019, Volume 1*, SciTePress, 2019, pp. 407–414.

[120] F. Rupp, B. Schnabel and K. Eckert, Easy and Complex: New Perspectives for Metadata Modeling Using RDF-Star and Named Graphs, in: *Knowledge Graphs and Semantic Web - 4th Iberoamerican Conference and third Indo-American Conference, KGSWC 2022, Madrid, Spain, November 21-23, 2022, Proceedings*, Communications in Computer and Information Science, Vol. 1686, Springer, 2022, pp. 246–262.

[121] M.S. Schlichtkrull, T.N. Kipf, P. Bloem, R. van den Berg, I. Titov and M. Welling, Modeling Relational Data with Graph Convolutional Networks, in: *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, Lecture Notes in Computer Science, Vol. 10843, Springer, 2018, pp. 593–607.

[122] A.T. Schreiber, G. Schreiber, H. Akkermans, A. Anjewierden, N. Shadbolt, R. de Hoog, W. Van de Velde and B. Wielinga, *Knowledge engineering and management: the CommonKADS methodology*, MIT press, 2000.

[123] J. Sequeda and O. Lassila, *Designing and Building Enterprise Knowledge Graphs*, Synthesis Lectures on Data, Semantics, and Knowledge, Morgan & Claypool Publishers, 2021.

[124] K. Shenoy, F. Ilievski, D. Garijo, D. Schwabe and P.A. Szekely, A study of the quality of Wikidata, *Journal of Web Semantics* **72** (2022), 100679.

[125] X. Shi and Y. Xiao, Modeling multi-mapping relations for precise cross-lingual entity alignment, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 813–822.

[126] R. Speer, J. Chin and C. Havasi, ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, *CoRR* **abs/1612.03975** (2016).

[127] G. Stoilos, G.B. Stamou, V. Tzouvaras, J.Z. Pan and I. Horrocks, Fuzzy OWL: Uncertainty and the Semantic Web, in: *Proceedings of the OWLED*05 Workshop on OWL: Experiences and Directions, Galway, Ireland, November 11-12, 2005*, CEUR Workshop Proceedings, Vol. 188, CEUR-WS.org, 2005.

[128] F.M. Suchanek, G. Kasneci and G. Weikum, Yago: a core of semantic knowledge, in: *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, ACM, 2007, pp. 697–706.

[129] Z. Sun, W. Hu and C. Li, Cross-lingual entity alignment via joint attribute-preserving embedding, in: *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I 16*, Springer, 2017, pp. 628–644.

[130] Z. Sun, W. Hu, Q. Zhang and Y. Qu, Bootstrapping entity alignment with knowledge graph embedding., in: *IJCAI*, Vol. 18, 2018.

[131] Z. Sun, Q. Zhang, W. Hu, C. Wang, M. Chen, F. Akrami and C. Li, A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs, *Proc. VLDB Endow.* **13**(11) (2020), 2326–2340.

[132] Z. Sun, Q. Zhang, W. Hu, C. Wang, M. Chen, F. Akrami and C. Li, A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs, *Proc. VLDB Endow.* **13**(11) (2020), 2326–2340.

[133] Z. Sun, Z. Deng, J. Nie and J. Tang, RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space, in: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

[134] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N.T. Doncheva, A. Roth, P. Bork, L.J. Jensen and C. von Mering, The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, *Nucleic Acids Res.* **45**(Database–Issue) (2017), D362–D368.

[135] X. Tang, J. Zhang, B. Chen, Y. Yang, H. Chen and C. Li, BERT-INT: a BERT-based interaction model for knowledge graph alignment, *interactions* **100** (2020), e1.

[136] B.D. Trisedya, J. Qi and R. Zhang, Entity alignment between knowledge graphs using attribute embeddings, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, 2019, pp. 297–304.

[137] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier and G. Bouchard, Complex Embeddings for Simple Link Prediction, in: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, JMLR Workshop and Conference Proceedings, Vol. 48, JMLR.org, 2016, pp. 2071–2080.

[138] D. Vrandecic and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85.

[139] W.E. Walker, P. Harremoës, J. Rotmans, J.P. Van Der Sluijs, M.B. Van Asselt, P. Janssen and M.P. Krayer von Krauss, Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support, *Integrated assessment* **4**(1) (2003), 5–17.

[140] M. Wan, X. Chen, L.M. Kaplan, J. Han, J. Gao and B. Zhao, From Truth Discovery to Trustworthy Opinion Discovery: An Uncertainty-Aware Quantitative Modeling Approach, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, B. Krishnapuram, M. Shah, A.J. Smola, C.C. Aggarwal, D. Shen and R. Rastogi, eds, ACM, 2016, pp. 1885–1894.

[141] J. Wang, K. Nie, X. Chen and J. Lei, SUKE: Embedding Model for Prediction in Uncertain Knowledge Graph, *IEEE Access* **9** (2021), 3871–3879.

[142] J. Wang, T. Wu and J. Zhang, Incorporating Uncertainty of Entities and Relations into Few-Shot Uncertain Knowledge Graph Embedding, in: *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy - 7th China Conference, CCKS 2022, Qinhuangdao, China, August 24-27, 2022, Revised Selected Papers*, Communications in Computer and Information Science, Vol. 1669, Springer, 2022, pp. 16–28.

[143] X. Wang, L. Chen, T. Ban, M. Usman, Y. Guan, S. Liu, T. Wu and H. Chen, Knowledge graph quality control: A survey, *Fundamental Research* **1**(5) (2021), 607–626. doi:https://doi.org/10.1016/j.fmre.2021.09.003.
[144] Z. Wang, Q. Lv, X. Lan and Y. Zhang, Cross-lingual knowledge graph alignment via graph convolutional networks, in: *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 349–357.
[145] G. Weikum, X.L. Dong, S. Razniewski and F.M. Suchanek, Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases, *Found. Trends Databases* **10**(2–4) (2021), 108–490.
[146] M.L. Wick, S. Singh, A. Kobren and A. McCallum, Assessing confidence of knowledge base content with an experimental study in entity resolution, in: *Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013*, ACM, 2013, pp. 13–18.
[147] W. Wu, H. Li, H. Wang and K.Q. Zhu, Probase: a probabilistic taxonomy for text understanding, in: *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, ACM, 2012, pp. 481–492.
[148] X. Wu, J. Wu, X. Fu, J. Li, P. Zhou and X. Jiang, Automatic Knowledge Graph Construction: A Report on the 2019 ICDM/ICBK Contest, in: *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, IEEE, 2019, pp. 1540–1545.
[149] Y. Wu, X. Liu, Y. Feng, Z. Wang and D. Zhao, Jointly Learning Entity and Relation Representations for Entity Alignment (2019), 240–249.
[150] Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan and D. Zhao, Relation-aware entity alignment for heterogeneous knowledge graphs, *arXiv preprint arXiv:1908.08210* (2019).
[151] R. Xie, Z. Liu, F. Lin and L. Lin, Does William Shakespeare REALLY Write Hamlet? Knowledge Representation Learning With Confidence, in: *Proceedings of the AAAI conference on artificial intelligence*, AAAI Press, 2018, pp. 4954–4961.
[152] K. Xu, L. Wang, M. Yu, Y. Feng, Y. Song, Z. Wang and D. Yu, Cross-lingual knowledge graph alignment via graph matching neural network, *arXiv preprint arXiv:1905.11605* (2019).
[153] B. Xue and L. Zou, Knowledge Graph Quality Management: A Comprehensive Survey, *IEEE Trans. Knowl. Data Eng.* **35**(5) (2023), 4969–4988.
[154] B. Yang, W. Yih, X. He, J. Gao and L. Deng, Embedding Entities and Relations for Learning and Inference in Knowledge Bases, in: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
[155] H. Yang, Y. Zou, P. Shi, W. Lu, J. Lin and X. Sun, Aligning Cross-Lingual Entities with Multi-Aspect Information, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, K. Inui, J. Jiang, V. Ng and X. Wan, eds, Association for Computational Linguistics, 2019, pp. 4430–4440.
[156] Y. Yang and J. Calmet, OntoBayes: An Ontology-Driven Uncertainty Model, in: *2005 International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA 2005), International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC 2005), 28-30 November 2005, Vienna, Austria*, IEEE Computer Society, 2005, pp. 457–463.
[157] A. Yates, M. Banko, M. Broadhead, M.J. Cafarella, O. Etzioni and S. Soderland, TextRunner: Open Information Extraction on the Web, in: *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, The Association for Computational Linguistics, 2007, pp. 25–26.
[158] R. Ye, X. Li, Y. Fang, H. Zang and M. Wang, A vectorized relational graph convolutional network for multi-relational network alignment., in: *IJCAI*, 2019, pp. 4135–4141.
[159] X. Yin, J. Han and P.S. Yu, Truth discovery with multiple conflicting information providers on the web, in: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, ACM, 2007, pp. 1048–1052.
[160] L.A. Zadeh, Toward a generalized theory of uncertainty (GTU)–an outline, *Inf. Sci.* **172**(1–2) (2005), 1–40.
[161] J. Zhang, T. Wu and G. Qi, Gaussian Metric Learning for Few-Shot Uncertain Knowledge Graph Completion, in: *Database Systems for Advanced Applications - 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11-14, 2021, Proceedings, Part I*, Lecture Notes in Computer Science, Vol. 12681, Springer, 2021, pp. 256–271.
[162] Q. Zhang, Z. Sun, W. Hu, M. Chen, L. Guo and Y. Qu, Multi-view knowledge graph embedding for entity alignment, *arXiv preprint arXiv:1906.02390* (2019).
[163] B. Zhao and J. Han, A probabilistic model for estimating real-valued truth from conflicting sources, *Proc. of QDB* **1817** (2012).
[164] B. Zhao, B.I.P. Rubinstein, J. Gemmell and J. Han, A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration, *Proc. VLDB Endow.* **5**(6) (2012), 550–561.
[165] Y. Zhao, J. Hou, Z. Yu, Y. Zhang and Q. Li, Confidence-Aware Embedding for Knowledge Graph Entity Typing, *Complex.* **2021** (2021), 3473849:1–3473849:8.
[166] Y. Zheng, G. Li and R. Cheng, DOCS: Domain-Aware Crowdsourcing System, *Proc. VLDB Endow.* **10**(4) (2016), 361–372.
[167] L. Zhong, J. Wu, Q. Li, H. Peng and X. Wu, A Comprehensive Survey on Automatic Knowledge Graph Construction, *ACM Comput. Surv.* **56**(4) (2024), 94:1–94:62.
[168] H. Zhu, R. Xie, Z. Liu and M. Sun, Iterative Entity Alignment via Joint Knowledge Embeddings., in: *IJCAI*, Vol. 17, 2017, pp. 4258–4264.
[169] Q. Zhu, X. Zhou, J. Wu, J. Tan and L. Guo, Neighborhood-Aware Attentional Representation for Multilingual Knowledge Graphs., in: *IJCAI*, 2019, pp. 1943–1949.
[170] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen and N. Zhang, LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities, *arXiv preprint arXiv:2305.13168* (2023).