Uncertainty Management in the Construction of Knowledge Graphs: A Survey

Lucas Jarnac[1]
Orange Research, Belfort, France
Université de Lorraine, CNRS, LORIA, Nancy, France

Yoan Chabot
Orange Research, Belfort, France

Miguel Couceiro
Université de Lorraine, CNRS, LORIA, Nancy, France
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. However, 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 represent 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. We then describe different knowledge extraction methods and 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.

2012 ACM Subject Classification Information systems → Information integration; Information systems → Uncertainty; Information systems → Graph-based database models

Keywords and phrases Knowledge reconciliation, Uncertainty, Heterogeneous sources, Knowledge graph construction

Digital Object Identifier 10.4230/TGDK.3.1.3

Category Survey

Received 2024-07-19 **Accepted** 2025-04-25 **Published** 2025-06-20

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 leading to conflicts. These are the effects of incompleteness, vagueness, fuzziness, invalidity, ambiguity, and timeliness, leading to uncertainty about the correctness of the data [33]. The uncertainty can be due to the source of the data (*e.g.,* a document written by an expert versus a non-expert

[1] Corresponding author

(a) French Wikipedia page of Jacques Chirac. **(b)** Wikidata page of Jacques Chirac.

Figure 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.

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). For example, on the French Wikipedia page of the former president of France Jacques Chirac[2], we can read that he was the mayor of Paris from March 25, 1977 to May 16, 1995 while on Wikidata[3] 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 stable over time [133]. Some facts are known to change including all facts that involve a period of time for which the fact is valid (*e.g.,* the mandate of a president or the place of residence of a person) or knowledge in specific domains may change regularly. For instance, in paleontology where excavations reveal new discoveries and modify knowledge previously established.

This raises questions about the meaning of data and knowledge. Data are uninterpretable signals (*e.g.,* numbers or characters). Information is data equipped with a meaning. In [138], 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" [69]. Many knowledge graphs (KGs) have been built to represent such data in recent years, and they have become a major asset for organizations, since KGs can support various downstream tasks such as knowledge and vocabulary indexing, as well as other applications in recommendation systems, question-answering systems, knowledge management, or search engine systems [68, 122]. 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 [100], 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 in 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 [81, 163, 93].

Uncertainty can also be found in knowledge and we can distinguish two types of knowledge: *objective knowledge* where 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* where multiple values can be accepted depending on their context and point of view (*e.g.,* the number of participants in

[2] `https://fr.wikipedia.org/wiki/Jacques_Chirac`
[3] `https://www.wikidata.org/wiki/Q2105`

a protest depending on the counting technique). Most KG construction methods do not take into account noisy facts and the uncertainty inherent in the 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 accurate as possible [122]. In this survey, we review different approaches to integrate knowledge uncertainty in the main steps of KG construction with current knowledge fusion methods and its representation in the graph. [127] surveys approaches and evaluation methods for KG refinement, particularly KG completion and error detection methods. In [98], the authors review truth discovery methods used in knowledge fusion before 2015. However, to the best of our knowledge, no survey specifically addresses uncertainty handling in KG construction.

The remainder of this survey is structured as follows. In Section 2, we describe our research methodology which allowed 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 for 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 surveys methods to construct a KG from uncertain knowledge. In this section, we present our research methodology for finding and selecting papers for 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 were 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 first looked 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. This survey covers the works addressing the representation of uncertainty in a KG published after 2004, while the representation of uncertainty in a vector space, which is a more recent research topic, covers those published since 2016. The distribution of publication years for data fusion methods presented in this survey is rather spread out starting from 2007. This is due to new models based on deep learning that are being explored to tackle these tasks.

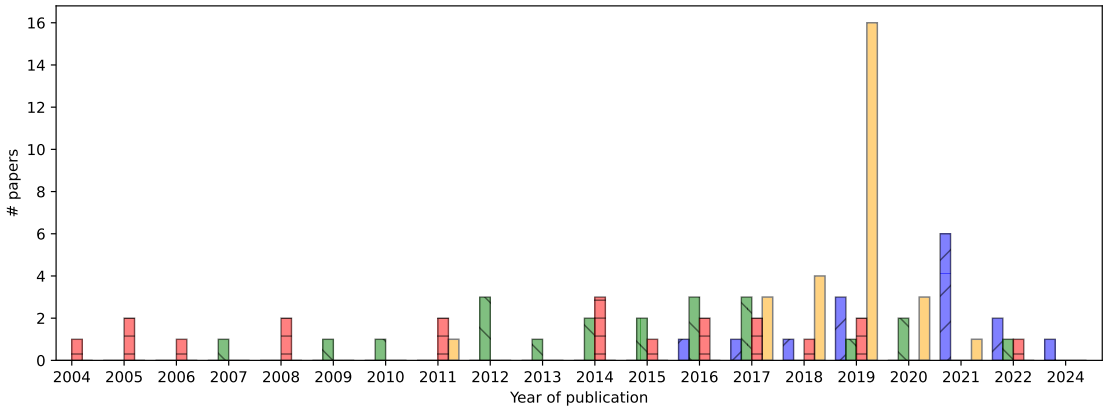

Figure 2 Distribution of paper publication years according to uncertain KG embedding ( ), knowledge fusion ( ), knowledge alignment ( ), uncertainty representation ( ).

3 Knowledge Graphs

Before going into further detail on reconciliation approaches, it is important to define KGs, which form 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 representation and abstraction of data, making them an ideal tool to manage knowledge of organizations in a sustainable way or to support search and querying applications [69], which led several companies to build their own KGs [122].

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 3.2 and Section 3.3.

3.1 What Is a Knowledge Graph?

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$\rangle$, which 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 [69, 122]. 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 by a schema or otherwise named an ontology, which itself can be represented as a graph embedded in the KG [42, 36]. To build a schema or ontology, different vocabularies or languages are available, such as RDF Schema[4] (RDFS) or OWL (Ontology Web Language) and OWL 2 [52]. RDFS is designed to describe RDF vocabularies specific to a domain by defining classes, subclass relationships, properties, domain, range and sub-property relationships. OWL is an ontology language and an extension of RDFS defined by Description Logics (DLs), which are a family of formal knowledge representation that enable reasoning about ontologies. Specifically, OWL DL is based on the description logic $\mathcal{SHOIN}$ [52] and OWL 2 DL is based on $\mathcal{SROIQ}$ [70]. OWL and OWL 2 allow the creation of ontologies with more complex constraints than RDFS (*e.g.,* cardinality restrictions). They include several sub-languages, each offering different levels of expressiveness depending on

[4] `https://www.w3.org/TR/rdf12-schema/`

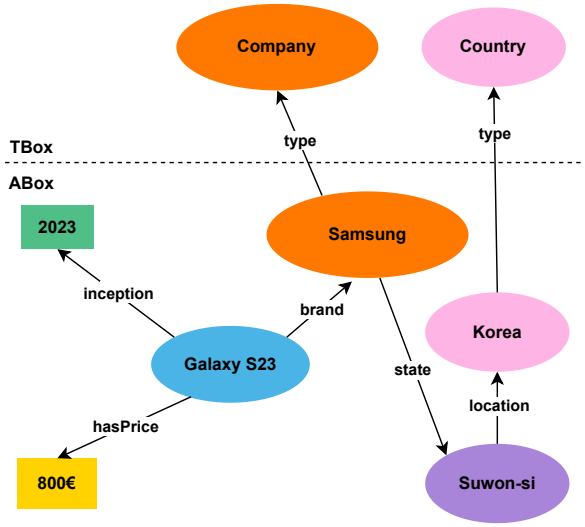

Figure 3 TBox stands for terminology box that contains classes and properties; ABox stands for assertion box that contains instances (*e.g.,* Galaxy S23) and values (*e.g.,* "2023").

the constructors applied (*e.g.,* disjunction, inverse roles, role transitivity, etc.) in the description logic that defines the language. In a KG with an ontology defined using a language based on DLs, two boxes coexist, namely the Terminology Box (TBox) and the Assertion Box (ABox). The TBox defines classes and properties and the 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 semi-structured representation of data, defined by its ontology, provides a clear and flexible semantic representation whose classes and relations can be easily added and connect a large number of domains [36].

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**[5], 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 [155]. 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, and references to Wikipedia articles. To maintain the quality of data, some constraints such as property or unique value constraints alert the user in case of suspicious input (*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 [162].

NELL is an intelligent computer agent that ran continuously between January 2010 and September 2018 according to the official NELL website[6] and that every day extracted knowledge from texts, tables, and lists on the web to feed a Knowledge Base (KB) [19]. To maintain the consistency of the KB, the knowledge integrator of NELL exploits relationships between predicates by respecting mutual exclusion (*e.g.,* an instance of a Human class cannot be an instance of

[5] https://www.wikidata.org/wiki/Wikidata:Main_Page
[6] http://rtw.ml.cmu.edu/rtw/

a Car class, since Human and Car are mutually exclusive) and type checking information. In addition, NELL components provide a probability for each candidate and a summary of the source supporting it.

YAGO is an ontology built on statements of Wikipedia that combines high coverage with high quality [145]. 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 KB built by extracting structured information from Wikipedia (*e.g.,* infoboxes) and making this information available on the Web [7]. Since DBpedia is populated from Wikipedia pages in different languages, the data retrieved can sometimes be conflicting. To manage these conflicts, DBpedia has a module called *Sieve* which performs quality assessments by computing some metrics such as the "recency" or the "reputation" of data before applying a fusion step based on these dimensions [111, 16].

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 users to complete and semi-automatically reconcile data 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 such as Google Search, Google Maps, and Google KG, nowadays, Freebase is closed, and its knowledge has been transferred to Wikidata [11].

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 associated with numerical values that intend to represent the level of uncertainty about the relation [142].

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 [51, 122, 139]. This led companies such as Google, Microsoft, Amazon, Facebook, Orange and IBM, to build their own KGs [122, 77].

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 [122], it contains general information about the world such as people, or places and allows users to take actions such as 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 [117] provides various 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.

Also, Orange bootstraps its KG from a set of terms of interest from a enterprise repository [77]. These terms of interest are aligned with their equivalent Wikidata entities. Then, an expansion is performed by retrieving their N-hops neighborhood to identify additional entities of interest. 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 [77, 76].

In [122], 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, which is the first step in the knowledge integration process. In this section, we present what knowledge extraction is and some well-known automatic approaches 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). These approaches inherently introduce uncertainty due to their imperfect accuracy in extracting facts. Finally, the definition of KG quality and related metrics is provided in Section 4.2.

4.1 Knowledge Extraction

Methods for populating a KG depend on the knowledge domain and the desired graph coverage. For example, one method could rely on the knowledge of domain experts and populate the graph manually (*e.g.,* by crowdsourcing, such as Wikidata [155] or Freebase [11]). However, this is a time-consuming process, especially if the graph is intended to be large and may suffer from quality issues [15, 140]. 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 rely on automatic construction methods [122]. In the following sections, we present the different tasks involved in knowledge extraction from various data sources.

4.1.1 From Texts

For a long time, the majority of data has been represented and exchanged in the form of text [106, 125]. Texts in all their forms (e.g. reports, articles, or any other textual documents) are an invaluable source of information, as they are the most widely used data formats in the world (*e.g.,* in the scientific research domain, where knowledge is communicated through scientific articles [74]). To leverage knowledge from texts as data sources to enrich a KG, we rely on a 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 [120]. The aim of information extraction is to identify entities, their attributes, and their relationships with other entities in text [165]. In general, this task is divided 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 [184, 58].

There are two main IE approaches in the literature [120]: 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 [120].

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 [174, 120]: automation, text heterogeneity and scalability. Automation means that the information extraction system must rely on unsupervised extraction strategies. Heterogeneity stands for the different domains and genres of text *e.g.,* a scholar journal versus a popular science journal. Since the extractions are performed in an unsupervised manner *i.e.,* without any labeling or predefined schema to support the extraction, this 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 common approaches to address these issues consist of pipelines composed of methods based on Natural Language Processing (NLP) [165].

One of the earliest examples of a traditional IE system is **KnowItAll** [43], which automates the domain-independent extraction of large collections of facts (*i.e.,* triples) from the Web. However, it is supported by an extensible ontology and a minimal set of generic rules for extracting entities and relations contained in its ontology. KnowItAll consists of 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 ..." → "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 [46], whereas Open IE no longer relies on predefined patterns and allows new information to be explored [120].

For example, **TextRunner** [174] which introduced the concept of Open IE, extracts a set of relational tuples without requiring human input. 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 which groups synonymous entities and relations together, since TextRunner has no predefined relations to guide extractions.

A slightly more recent approach is **ReVerb** [46]. 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 on relational sentences are introduced: a syntactic constraint

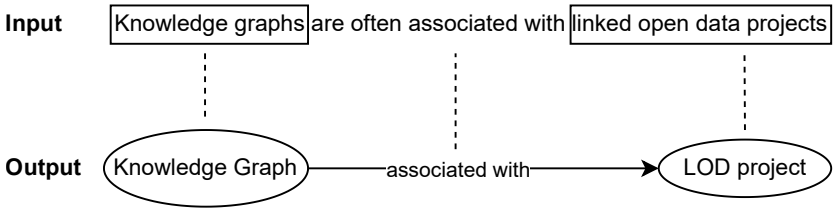

Figure 4 Illustration of knowledge extraction from a single sentence. LOD stands for Linked Open Data.

and a lexical constraint. First, 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 1,000 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 5.3).

OLLIE [108] extends 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 factual. 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 [75], 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 [75].

4.1.2 From the Web

The Web contains a 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, partly 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 [112] is an agent that takes an initial ontology consisting of categories and relations, which is used to define learning tasks such as category classification, relation classification, or entity resolution. 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. NELL reads facts from the Web and incrementally refines its KB by removing incorrect ones from a set of labeled data and user feedback on the trustworthiness of the extracted facts. The extracted facts are then stored in the 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 containing 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. Fact extraction is performed using four different extractors from: text documents, HTML trees, HTML tables, and human annotated pages. To merge the extractors they construct a feature vector for each extracted triple and apply a binary classifier to compute the probability of the fact being true given the feature vector. Each predicate is associated with a different classifier. The feature vector contains, for each extractor, the square root of the number of sources from which the triple was extracted and the mean score

of the extractions across these sources. They assume that the confidence scores of 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.

Probase [164] does not consider knowledge extracted from the Web to be deterministic but rather 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 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 using an iterative learning algorithm that extracts pairs (x, y) that verify an *isA* relation between x and y, and 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 pieces of evidence 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 models and Large Language Models (LLMs), some triple extraction tasks are now successfully carried out by such models. [114] 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 [125]. For example, a workshop on KB construction from pre-trained language models (KBC-LM[7]) and a challenge on language models for KB construction (LM-KBC[8]) are now proposed at the International Semantic Web Conference (ISWC[9]).

In [58], the authors use the BERT model for NER and RE tasks to build a biomedical KG. In [57], the authors exploit knowledge encoded in LLM parameters (*a.k.a.* parametric knowledge [125]) to feed a KG by harvesting 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 shots 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.

However, Pan *et al.* [125] 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 they are mainly applied to generic domains and perform poorly on specific domains. They also lack accuracy with numerical facts (*e.g.,* the birthday of a person) and have difficulty retaining knowledge related to long-tail entities. In addition, LLMs are subject to various biases (*e.g.,* gender bias) that are inherent in the training data. Finally, LLMs do not provide any provenance or reliability information for the extracted knowledge [125], which

[7] https://lm-kbc.github.io/workshop2024/
[8] https://lm-kbc.github.io/challenge2024/
[9] https://iswc2024.semanticweb.org/

can be an obstacle for many knowledge fusion approaches presented in Section 8. In [187], the authors evaluate the ability of LLMs, particularly different GPT models, on KG construction and reasoning tasks (*i.e.,* link prediction and question answering) under zero-shot and one-shot settings. The authors also point out that LLMs do not outperform state-of-the-art models for KG construction and have limitations in recognizing long-tail knowledge.

4.2 Quality and Metrics

Assessing the quality of the constructed KG is important since it is practically impossible to obtain a perfect KG, especially when it 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 [160, 170, 69, 67]: completeness, accuracy, timeliness, availability, and redundancy.

Completeness refers to the coverage of knowledge within the specific domain the KG is intended to represent. The evaluation of this dimension depends on the assumption made about the KG [50]: the Open World Assumption (OWA), the Closed World Assumption (CWA), and the Local Closed World Assumption (LCWA). Under the OWA, if a triple is not present in the KG, it is not necessarily incorrect but rather unknown. Conversely, under the CWA if a triple is not present in the KG, then it is considered incorrect. Finally, under the LCWA if the KG knows at least one object or value v for a predicate p associated with the subject s, it is assumed to know all the values of the pair (s, p). For example, this dimension can be measured as the number of instances represented in the KG relative to the total expected number of instances. Some studies have focused on estimating the expected total number. More details on these measures can be found in [160].

Accuracy corresponds to the correctness of the facts in the KG. In [162], 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. For example, this dimension can be measured by the response time to queries or by the level of accessibility to KG data (*e.g.,* RDF, Turtle, JSON-LD, or SPARQL query service).

Redundancy assesses whether different statements express the same fact, which may require an entity resolution task where duplicates are aligned and then the triples associated with these duplicates are fused.

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 data selection issues in terms of both source and quality. The metadata can also support knowledge fusion approaches by serving as prior knowledge, as we describe in Section 8. Other metrics are proposed in the survey [160] for each quality dimension.

5 Knowledge Graph Refinement

A KG can be populated by human effort, or by automatic knowlegde extraction approaches from heterogeneous sources (*e.g.,* tables, texts, databases, etc.) as presented in Section 4. The advantage of using multiple sources is twofold: ensuring knowledge coverage and identifying inconsistencies by leveraging collective wisdom [41]. However, the world is uncertain and data sources are of varying quality leading to uncertain knowledge, which must be handled in the integration process *w.r.t.* the quality dimensions listed in Section 4.2. The causes of uncertainty are presented in Section 5.1. We provide a brief overview of several methods for integrating data into a KB under uncertainty in Section 5.2. Section 5.3 presents our theoretical data integration pipeline, designed to address knowledge uncertainty and enrich the KG.

5.1 Knowledge Deltas

Uncertainty is everywhere in knowledge and can take the form of invalidity, vagueness, fuzziness, timeliness, ambiguity, and incompleteness according to [34]. 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 [156]. The possible causes of uncertainty are [1, 156]: *(i)* a lack of knowledge; *(ii)* a semantic mismatch or a lack of semantic precision, and *(iii)* a lack of machine precision.

When a KG is constructed from multiple heterogeneous data sources, uncertainty can lead to the emergence of knowledge deltas, characterized by differences in the information or facts they contain. These knowledge deltas may occur between two data sources on the same subject, *e.g.,* differences in specificity and contradictions. For example, a very specific data source A and a generic data source B may provide information on the same topic but with different terms, increasing the risk of knowledge delta. It is also possible for a data source to contradict itself, one possible way to detect these deltas is to compare the data source with itself by "reflecting on data patterns or extrapolation to complete missing information and/or detect wrong ones" according to [33]. On the other hand, duplicates can also occur when the two data sources provide exactly the same knowledge, which needs to be managed for reasons of scalability and KG quality.

We use some examples to illustrate the various forms that uncertainty can take. 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 provides invalid information: the date of renaming of the Paris region to "Île-de-France" is invalid in the Wikipedia page[10];
- **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[11] for the "located in the administrative territorial entity" property, which contains additional information such as the day, month, and year;

[10] https://en.wikipedia.org/w/index.php?title=Paris&oldid=1197869134
[11] https://www.wikidata.org/w/index.php?title=Q90&oldid=2058313448

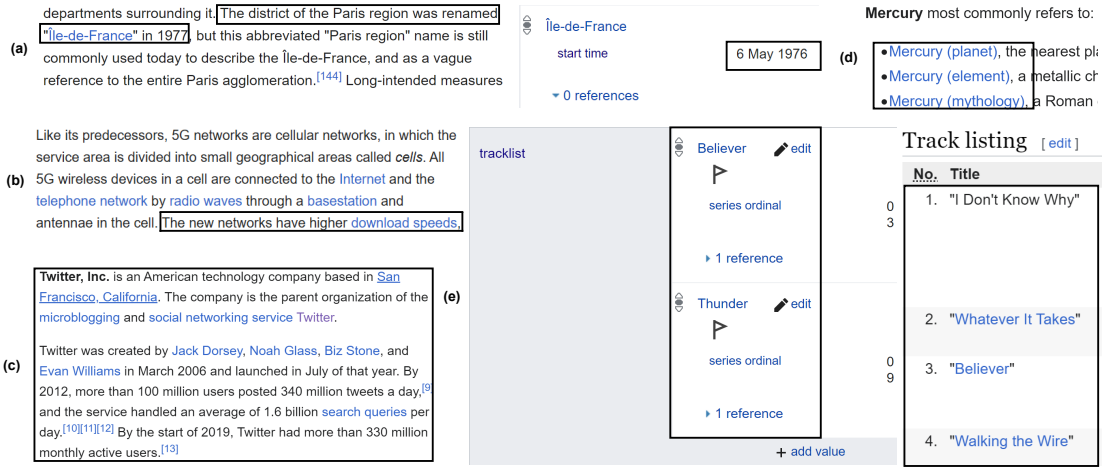

Figure 5 Illustration of the different possible deltas about some topics between English Wikipedia and Wikidata: (a) invalidity + vagueness, (b) fuzziness, (c) timeliness, (d) ambiguity, and (e) incompleteness.

- **Fuzziness:** t states a fuzzy truth, where the range of values is itself imprecise. If we focus only on the sentence within the black box in Figure 5 (b) of the Wikipedia article[12], it indicates that the 5G network has a higher peak download speed, but without specified lower and upper bounds.
- **Timeliness:** a data source may provide the statement t which is no longer valid at the current time, unlike another source, which may provide an updated version of t. As in Figure 5 (c), on the Wikipedia page[13] of May 10 2022, "Twitter" had not yet been renamed "X". This information has now been changed, otherwise there would have been an update issue;
- **Ambiguity:** t has multiple interpretations. As shown in Figure 5 (d), Mercury[14] can be a planet, an element, or a god in mythology;
- **Incompleteness:** t gives incomplete information. As in Figure 5 (e), the tracklist of the album "Evolve" by the group Imagine Dragons on Wikidata[15] contains fewer songs than in Wikipedia[16].

The appearance 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 an outdated data source. 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. 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 types of deltas leading to conflicts into two classes as depicted in Figure 6, namely *Specificity* that stands for a difference between two data sources in the specificity of knowledge and *Contradictory* that stands for an incompatibility of knowledge. We classify *Fuzziness*, *Incompleteness*, and *Vagueness* deltas in the specificity category. These deltas lead to different levels of specificity between the knowledge of two data sources. This knowledge is not necessarily wrong, 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 wrong.

[12] https://en.wikipedia.org/wiki/5G
[13] https://en.wikipedia.org/w/index.php?title=Twitter,_Inc.&oldid=1087087372
[14] https://en.wikipedia.org/wiki/Mercury
[15] https://www.wikidata.org/w/index.php?title=Q29868187&oldid=2009666363
[16] https://en.wikipedia.org/w/index.php?title=Evolve_(Imagine_Dragons_album)&oldid=1197244329

In [9], the authors distinguish two types of data conflicts from a data fusion perspective: *contradictions* and *uncertainties*. The authors define *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 adopt a different definition of *uncertainty*. We define the second type of conflict as a difference in the specificity of knowledge, as illustrated in Figure 6. In this survey, "uncertainty" is a more general term whose sources lie in knowledge deltas and the inaccuracy of each step in the knowledge integration pipeline, including knowledge acquisition.

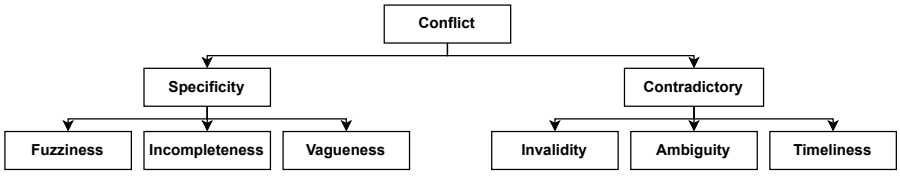

Figure 6 Our proposed classification of knowledge deltas by type of resulting conflict (difference in specificity or contradiction).

In [5, 177], 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 quantification, and the integration of uncertainty in the KG. In such a view, the uncertainty should be considered everywhere in the data integration pipeline, including its representation in the KG. In Section 5.3, we present our ideal data integration pipeline that addresses the aforementioned requirements.

5.2 Integrating Data Under Uncertainty

Integrating data from multiple sources can introduce inconsistencies and uncertainty into a database. Bleiholder and Naumann [9] describe the data integration process in three steps: 1) schema mapping, 2) duplicate detection, and 3) data merging. The first step establishes a common schema among data sources, the second step aligns duplicates and detects inconsistent representations for the same entity, and the final step combines and resolves the various inconsistencies (e.g., contradictions) to produce a unified representation. In [39], the authors mention that uncertainty can arise for several reasons, such as the approximation of semantic mappings between data sources and the mediated schema (*i.e.,* the integration schema grouping all sources), the extraction techniques used to extract data from unstructured sources, and also mention uncertainty at the application level when querying with the transformation of keywords into a set of candidate structured queries. In [39, 136], the authors address the schema mapping step, considering uncertainty through probabilistic schema mappings and how to answer queries on the mediated schema. Different strategies can be used to deal with inconsistencies. One approach involves blaming the most recent assertion that caused the inconsistency in the KB [107]. In [28], Amo et al. consider two methods. The first is to allow inconsistencies to remain in the KB and reason about them using paraconsistent logic [8], while the second approach seeks to get rid of inconsistencies to obtain a coherent KB. Belief revision addresses the latter approach by updating the knowledge in a KB when a contradiction is encountered. The update is done by revising the KB, where some beliefs must be retracted [56]. The formalism used to represent beliefs and the nature of the relationship between explicit and implicit beliefs play an important role in the belief revision

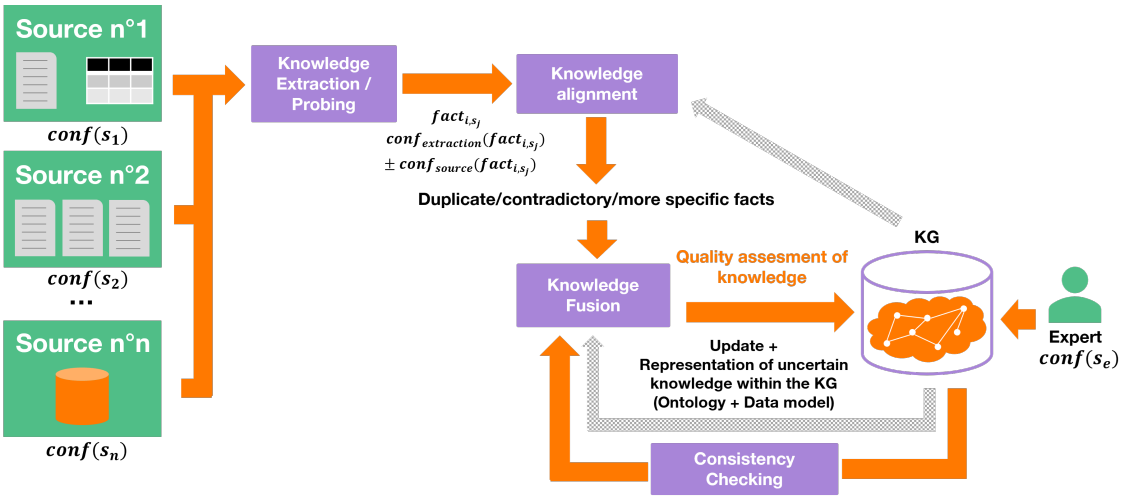

Figure 7 Process of integrating uncertain knowledge extracted from heterogeneous sources, ideally linked to different confidence scores. Three steps form knowledge integration: (1) alignment, (2) fusion, and (3) consistency checking.

process [56]. In [107], Martins and Shapiro propose a tailored logic for belief revision systems. Their system tracks the support for each proposition in the KB and applies inference rules of the logic to compute dependencies between propositions. These dependencies help identifying sources of inconsistency and guide the revision process.

5.3 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 humans cannot be experts in every domain, involving mistakes and extraction algorithms rarely achieve perfect precision. Errors can occur at various 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 [46] is associated with a confidence score obtained from a logistic regression. Another work [96] focuses on estimating a confidence score for the slot filling task, which consists of 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 the performance of the task and that the correctness of the values and the estimated confidence are strongly correlated. In [163], the authors estimate confidence scores for an entity alignment task that represent the marginal probability that a set of mentions all refer to the same entity. Therefore, there is a need to consider these confidence scores and represent them as triple metadata along with their provenance information.

From this perspective, we propose an ideal pipeline of data integration from heterogeneous sources depicted in Figure 7. After extraction, knowledge integration often involves the following two modules [10]: *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 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 $conf_{extract}(fact_{i,s_j})$, possibly 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 in specificity, and contradictions between the extracted facts and the KG. The next step *Knowledge Fusion* defines a policy to resolve conflicting facts and keep the information as consistent, specific, and complete as possible. In addition, it estimates the quality of both data sources and extracted facts by assigning them confidence scores. Then, the knowledge in the KG is updated with its confidence scores and their provenance information since a user may want to query the KG about the confidence of triples *w.r.t.* quality dimensions. The last step checks the consistency within the KG. Since this step is performed after the enrichment of the KG, and since this survey focuses on 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 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 the uncertainty in an embedding space have recently been investigated.

We describe the ideal data integration policy through an example. The integration pipeline takes as input a set of sources $\mathcal{S}$ containing a set of facts $\mathcal{F}$ and aims to construct a $\mathcal{KG}$ as output. Each fact $f_{\mathcal{S}} \in \mathcal{F}$ is represented as a triple of the form (e, r, v) where e is an entity, r is a relation, and v is either a literal or another entity. The alignment of each fact with $\mathcal{KG}$ consists in finding a correspondence between the pair (e, r) from the fact and an existing $(entity, relation)$ pair in the current state of $\mathcal{KG}$. If such a correspondence does not exist, the fact is added to $\mathcal{KG}$ along with its associated confidence score and provenance information. Otherwise, if such a correspondence exists, *i.e.,* $f_{\mathcal{S}}$ and $f_{\mathcal{KG}}$ share the same $(entity, relation)$ pair, the pipeline applies the following integration policy:

(1) If $f_{\mathcal{S}}$ is more specific than $f_{\mathcal{KG}}$, then $f_{\mathcal{KG}}$ is replaced by $f_{\mathcal{S}}$ and the confidence score of the source $\mathcal{S}$ is increased. If such comparisons are not possible, the strategy is to keep both facts;

(2) Otherwise, if $f_{\mathcal{S}}$ is a duplicate of $f_{\mathcal{KG}}$, only the provenance of $f_{\mathcal{S}}$ is added to the existing fact and the confidence scores of the sources providing it are increased;

(3) Otherwise, if $f_{\mathcal{S}}$ contradicts $f_{\mathcal{KG}}$, the integration pipeline resolves the conflict by finding the most trustworthy value. The confidence of the source providing the least trustworthy fact is decreased, while the confidence score of the source providing the most trustworthy one is increased.

This integration policy aligns with the fusion approaches presented in Section 8.2. These approaches follow the intuition that the reliability of a source increases when it provides facts that are estimated to be correct, and a fact is estimated to be correct if it is supported by reliable sources. The implementation of this policy relies on the following key elements:

- **Estimating specificity** cannot be performed globally across all knowledge domains, as it is a context-sensitive and relation-dependent task. A KG does not have a total order, but contextualized partial orders can be deduced from the KG taxonomy when available with relations such as *subclassOf*, *instanceOf*, or *partOf* and can help determine whether one value is more specific than another [6, 81]. For example, (*White house*, *location*, *United States*) versus (*White house*, *location*, *Washington*): the most specific fact, namely (*White house*, *location*, *Washington*) should be kept by the integration policy. In addition, it may be interesting to leverage the generic information *United States* by deviating from it such as (*White house*, *country*, *United States*) to enrich the graph.

- **Confidence scores** from the upstream steps of knowledge fusion must be calibrated to represent them as probabilities on the same scale (*e.g.* via Platt scaling or isotonic regression) [35] to make them comparable and initialize the fusion models summarized in Section 8. Only the confidence scores estimated within the fusion model itself, namely those representing the reliability of the sources and the trustworthiness of facts are updated during the fusion process.

Algorithm 1 illustrates our vision of the policy for integrating conflicting data. It provides a straightforward guideline and corresponds closely to the framework of knowledge fusion models presented in Section 8.

As mentioned in the process above, we need to maintain a history of the provenance of the facts, as provenance information is essential for the quality of the KG, but could also be used for future conflict resolution or KG updating [73]. For this purpose, there exists a normative ontology called PROV-O [89] that includes provenance information. 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.

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 with a confidence by an extraction algorithm $conf_{extract}(fact_{\mathcal{S}})$ 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
```

6 Uncertain Knowledge Graph Embedding

Embedding methods allow the KG to be represented in an n-dimensional vector space, *i.e.,* its entities and relations are n-vectors. These embeddings attempt to preserve the structural properties of the graph, making it easier to manipulate the graph for machine learning applications such as link prediction, completion, or node classification [78]. A wide range of embedding models have emerged, such as TransE [12], DistMult [171], ComplEx [154], RotatE [150], neural networks applied to graphs such as RGCN [137], and GCN [85]. Additionally, embeddings are also increasingly used in the construction of KGs, for example, for knowledge alignment [48], or other KG refinement tasks [68]. Most embedding approaches do not include knowledge uncertainty 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 introduce bias into the representation and impact further applications. Given the importance of embedding methods in both KG applications and construction, we believe it is useful to gather such methods that incorporate uncertainty, expressed as 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 define uncertain KGs as follows. An *uncertain knowledge graph* (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 associated with triples in the KG relies on the plausibility of the triples, but most 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 been studied only recently. Each of the following models has its own specific approach to incorporating a confidence score into their modeling. However, most models use a scoring function derived from existing deterministic KG embedding models, incorporating confidence scores, for instance, into the loss function used to train the embeddings.

UKGE [25] 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 both observed and unseen triples. It maps the scoring function results to confidence scores using two different mapping functions: a logistic function and the bounded rectifier function. For the tasks of 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 [158] argues that UKGE does not fully exploit the structural information of facts. To improve this, SUKE has two components: an evaluator and a confidence generator. The evaluator assigns a structural score and an uncertainty score to each fact, which are jointly used to define the rationality of a fact. This rationality score is then used to generate a set of candidate triples, which are then fed into the confidence generator. The generator outputs confidence scores for these triples based on the uncertainty score provided by the evaluator. The plausibility of facts is computed with the DistMult [171] 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.

Table 1 Uncertain embedding models taking into account at least one numerical value on edges with their scoring function, tasks handled, datasets on which they are evaluated, and the year of the publication.

Model	Scoring function	Evaluation task	Dataset	Year
IIKE [47]	TransE	Link prediction Triple classification	NELL FB15K (synthetic)	2016
URGE[71]	Matrix Factorization (proximity preservation)	Node Clustering Node Classification k-NN search	PPI DBLP	2017
CKRL [168]	TransE	KG Noise Detection KG Completion Triple Classification	FB15K (synthetic)	2018
GTransE [83]	TransE	KG Completion	FB15K-237 (synthetic) NELL (synthetic)	2019
UKGE [25]	DistMult	Confidence prediction Relation fact ranking Relation fact classification	CN15K NL27K PPI5K	2019
UOKGE [14]	TransE	Confidence Prediction Triples Classification	CN15K	2019
SUKE [158]	DistMult	Link prediction Fact classification	CN15K NL27K PPI5K	2021
BEURRE [24]	Gumbel boxes	Confidence prediction Relation fact ranking	CN15K NELL27k	2021
PASSLEAF [26]	RotatE ComplEx DistMult	Tail Entity Prediction Confidence Prediction	PPI5K NL27K CN15K WN18RR FB15K237	2021
GMUC [178]	Minimum Similarity	Link Prediction Confidence Prediction	NL27K	2021
ConfE [182]	RESCAL	Entity Type Noise Detection Entity Type Prediction	FB15kET (synthetic) YAGO43k (synthetic)	2021
FocusE [124]	Modifiable Scoring Layer	Link Prediction	O*NET20K CN15K NL27K PPI5K	2021
WaExt [86]	TransE TransH DistMult ComplEx	Link Prediction Triple Classification	CN15K NL27K PPI5K	2022
Wang *et al.* [159]	Similarity	Confidence Prediction Tail Entity Prediction	CN15K NL27K	2022
MUKGE [101]	Circular correlation	Confidence Prediction Relation Fact Ranking Relation Fact Classification	CN15K NL27K PPI5K	2024

BEURRE [24] 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 [83] embeds uncertainty in a translational model by extending the well-known TransE [12]. Uncertainty, represented by a confidence score, is incorporated 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 is a margin parameter and $[x]_{+}$ represents the positive part of x. The scoring function f corresponds to the L1 or L2-norm. Thus, with this loss function if a triple (h, r, t) has a high confidence score, it will tend to satisfy $h + r = t$, otherwise the entity t will tend to diverge from $h + r$. Before GTransE, the same authors introduced **CTransE [82]** which is a related model that omits the hyperparameter α used as an exponent of the confidence score.

IIKE [47] models confidence within the embedding space using 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 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 scores from knowledge extraction. They then apply stochastic gradient descent to refine the embeddings iteratively.

PASSLEAF [26] decomposes the model into two components: a confidence score prediction framework that adapts scoring function from existing models, *e.g.,* ComplEx [154] or RotatE [150], and a semi-supervised learning framework.

For the UKG completion task, each relation must have sufficient training examples to perform correctly. **GMUC [178]** tackles the few-shot UKG completion task for long-tail relations. GMUC learns a Gaussian similarity metric that enables the prediction of missing facts and their confidence scores from a limited number of training examples. The model encodes a support set, comprising a few facts along with their confidence scores, and a query into multidimensional Gaussian distributions. The query consists of *(head, relation)* pairs where the tail and confidence score must be predicted. A Gaussian matching function is then applied to generate a similarity distribution between the query and the support set. GMUC outperforms the UKGE model on link prediction and confidence prediction tasks on NL27K and three NL27K-derived datasets with added noise.

UOKGE [14] learns embeddings of uncertain ontology-aware KGs based on confidence scores. It encodes an instance $e_i \in E$ as a point represented by an n-dimensional vector, a class $c_i \in C$ as a sphere $s_i(c_i, \rho_i)$ where c_i denotes the center of the sphere and p_i its 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) defines the center of the sphere, with p_i^d representing the domain, p_i^r representing the range, and ϱ_i is the radius. It then introduces a mapping function to rescale values between 0 and 1 to represent uncertainty. Six distinct gap functions are defined to encode uncertainty for six types of relations: type, domain, range, subclass, sup-property, and remaining properties. The model then minimizes the mean squared error (MSE) between the confidence scores and the corresponding gap functions.

FocusE [124] improves KG embeddings with numerical values on edges by intervening 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 manner that maximizes the margin between the scores of true triples and their corruptions. Given the score function of an embedding model $f(t)$, where t is a triple, they use a nonlinear Softplus function σ such that the score provided by $f(t)$ is greater than or equal to zero:

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

The numerical value associated with an edge is then expressed through α as follows:

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

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

ConfE [182] encodes the tuples (e, τ), where e is an entity and τ is an entity type, by considering the uncertainty in each tuple. They treat entities and entity types as distinct elements in a KG and learn embeddings of entities and entity types in two separate 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. Uncertainty is incorporated into the loss function as follows:

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

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

CKRL [168] 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 as follows:

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

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

WaExt [86] embeds triples of a KG by incorporating the weight w associated with an edge (h, r, t) into the scoring function as follows:

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

and then minimizes a margin ranking loss function.

Wang *et al.* [159] 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 [101] aims to improve the generation of unseen facts for KGE training. The authors argue that PSL cannot leverage global multi-path information, leading to information loss when estimating the confidence of unseen facts. Indeed, PSL only considers information from simple logical rules with a path length of two, as used in UKGE [25] (*e.g.,* $(college, synonym, university) \wedge (university, synonym, institute) \rightarrow (college, synonym, institute)$), and does not consider other paths in the graph between the subject and object of the inferred relation. To address this issue, MUKGE introduces an algorithm called Uncertain ResourceRank, which infers confidence scores for unseen triples based on the relevance of the entity pairs (subject, object). The relevance of an entity pair is computed with respect to the directed paths between subject and object in the KG. MUKGE uses circular correlation as the scoring function, and applies either the sigmoid or the bounded rectifier function as the function to obtain the triple confidence. Then the authors design the loss function to align each positive triple to its corresponding confidence score. The

authors evaluate their model on confidence prediction, relation fact ranking, and relation fact classification. For these three tasks, MUKGE outperforms the BEURRE and UKGE models, particularly focusing on asymmetric relations. However, for all relation types, the performance is competitive with 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 derived from uncertain KBs are commonly used in UKG completion or confidence prediction tasks, as presented in the previous section [124]. **CN15K** is a subset of ConceptNet (discussed in Section 3.2) where the numerical values represent the uncertainty of the triples [142]. 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 using an Expectation Maximization (EM) algorithm and a semi-supervised learning method. **PPI5K** is a KG that represents the protein-protein interactions, where the numerical values indicate the confidence of the relations [151]. **O*NET20K** is a dataset introduced by [124], containing descriptions of 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 specific probability distributions.

7 Knowledge Alignment

In Section 4, we discussed enriching a KG through extraction methods, however, it is also possible to extend a KG by leveraging existing KGs. To achieve this, the two KGs must be aligned. In this section, we first introduce the task and mainly present embedding-based entity alignment approaches. Knowledge alignment, also known as knowledge resolution or knowledge matching, is the process of finding relationships or correspondences between entities of different ontologies [45]. It represents one of the steps in identifying candidate entities for knowledge fusion. For example, in Figure 8, the entity "Galaxy S23" in both graphs refers to the same real world entity but originates from two different sources. This task copes with the "redundancy" quality dimension (Section 4.2). Whether at the instance level or at the ontology level, many works tackle the knowledge alignment task. This section aims to provide a brief overview of the knowledge alignment task and approaches by collecting various existing surveys [48, 44, 148].

The authors of [45] distinguish different types of matching including semantic and syntactic approaches, such as string-based, language-based, subgraph-based, rule-based, embedding-based, or relational-based methods. An example of a rule-based method is presented in [79]. For each relation R_k across the two domains, the following rules are defined:

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

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

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

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 grouping entities. For example, Nguyen *et al.* [115] propose different strategies for blocking based on the description of entities, 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

Table 2 Recent embedding-based approaches that tackle knowledge alignment task.

Model	Embedding			Method	Learning
	One-hop	Multi-hop	Path		
MTransE [23]	•			Mapping	Supervised
IPTransE [185]			•	Sharing	Semi-supervised
JAPE [146]	•			Sharing	Supervised
BootEA [147]	•			Swapping	Semi-supervised
KDCoE [22]	•			Mapping	Semi-supervised
NTAM [95]	•			Swapping	Supervised
GCNAlign [161]		•		Mapping	Supervised
AttrE [153]	•			Sharing	Supervised
IMUSE [63]	•			Sharing	Supervised
SEA [90]	•			Mapping	Supervised
RSN4EA [55]			•	Sharing	Supervised
GMNN [169]		•		Swapping	Supervised
MuGNN [18]		•		Mapping	Supervised
OTEA [128]	•			Mapping	Supervised
NAEA [186]		•		Swapping	Supervised
AVR-GCN [175]		•		Swapping	Supervised
MultiKE [179]	•			Swapping	Supervised
RDGCN [166]		•		Mapping	Supervised
KECG [91]		•		Mapping	Supervised
HGCN [167]		•		Mapping	Supervised
MMEA [141]	•			Sharing	Supervised
HMAN [172]		•		Mapping	Supervised
AKE [99]	•			Mapping	Supervised
RREA [105]		•		Sharing	Supervised
BERT_INT [152]		•		Sharing	Supervised
MTransE+RotatE [149]	•			Sharing	Supervised

in the description of the URI (*e.g.,* URI infix) to create new blocks. After an optional blocking step, knowledge alignment methods are performed. Most recent KG alignment methods are deep learning methods based on graph embeddings. Among them, [48] distinguish three strategies: Sharing, Swapping, and Mapping. *Sharing* updates the entity embeddings produced by the embedding module according to the seed alignments. *Swapping* updates the entity embeddings produced by the embedding module according to the seed alignments 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 the 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 others the path in the graph is an important aspect. We provide Table 2, which is strongly inspired by [48, 148] and summarizes these different existing alignment approaches to give an overview.

The authors of [48] highlight that BERT_INT outperforms all models in terms of effectiveness and efficiency, especially when the KGs contain very similar factual information. In fact, the alignment models that use language models, such as BERT_INT, are the most efficient for this task. They also highlight the critical factors that affect the effectiveness of relation-based and attribute-based alignment methods, such as:

- the depth of neighbors considered;
- the strategy of negative sampling (for training), as the number of negatives are considered, the performances decrease;

- the input KGs to align, *e.g.,* for OpenEA datasets, it is not necessary to use attribute information; factual information is sufficient.

With the development of KGE models, KG alignment approaches based on embeddings have gained significant interest in recent years. However, probabilistic models also address this task. For example, in [144], the authors introduce a holistic model called PARIS, which aligns both instances and schemas of KGs. The model employs a probabilistic algorithm centered around the notion of "functionality". They define the functionality of a relation r as follows:

$$fun(r) = \frac{|\{x \mid \exists y,\, r(x,y)\}|}{|\{(x,y) \mid r(x,y)\}|}$$

Where x and y are instances. In this way, they incorporate the functionality of relations into the computation of probabilities for equivalences between instances. They also define the computations of the probabilities that a relation is a sub-relation of another based on the probabilities of equivalence between instances, and similarly for classes. Finally, to find equivalences they use an iterative approach that first computes the probabilities of equivalence between instances, then the probabilities for sub-relations, and the probabilities of equivalence between classes until the maximum assignments stabilize. PRASEMap [129, 130] is an unsupervised KG alignment approach that combines probabilistic reasoning (PR module) and semantic embeddings (SE module). The PR module uses the PARIS [144] model, while the SE module relies on GCNAlign [161] (other embedding models can be used). The PR module identifies cross-KG relations, literals, and entity mappings. The SE module leverages the entity mappings generated by the PR module as seed alignments for training and outputs additional entity mappings. Finally, the PR module is performed by incorporating embeddings and mappings from the SE module to perform more precise probabilistic reasoning. These last two steps can be iteratively performed to find more alignments. The integration of embeddings into the PR module is achieved by adding a weighted term that computes the similarity between two entities using the cosine similarity of their embeddings.

In the literature, handling uncertainty in KG alignment is often associated with the alignment of ontologies. In [1], the authors present and distinguish various operations on ontologies, such as mapping, merging, and alignment. They then describe methods for handling these operations under uncertainty, using probabilistic and logical models (such as Markov Logic Networks), fuzzy logic, Dempster-Shafer theory and hybrid methods combining machine learning with probabilistic approaches.

8 Uncertain Knowledge Fusion

In the previous section, we introduced the KG alignment task, along with a summary of the embedding-based approaches that tackle it. This task aims to identify equivalent entities and group them into clusters. Once these clusters have been formed, the next step consists in fusing the attributes of the entities within each cluster (as illustrated in Figure 7), as these attributes may be redundant, inconsistent, contradictory, or expressed at different levels of specificity. We first define the task in Section 8.1, and we present the various fusion approaches in Section 8.2.

8.1 Task Definition

The knowledge fusion step involves combining various pieces of information about the same entity or concept from multiple data sources into a consistent and a unified form, addressing the different deltas listed in Section 5.1 [67, 121]. The authors of [40] identify three broad goals to be achieved in this challenging task:

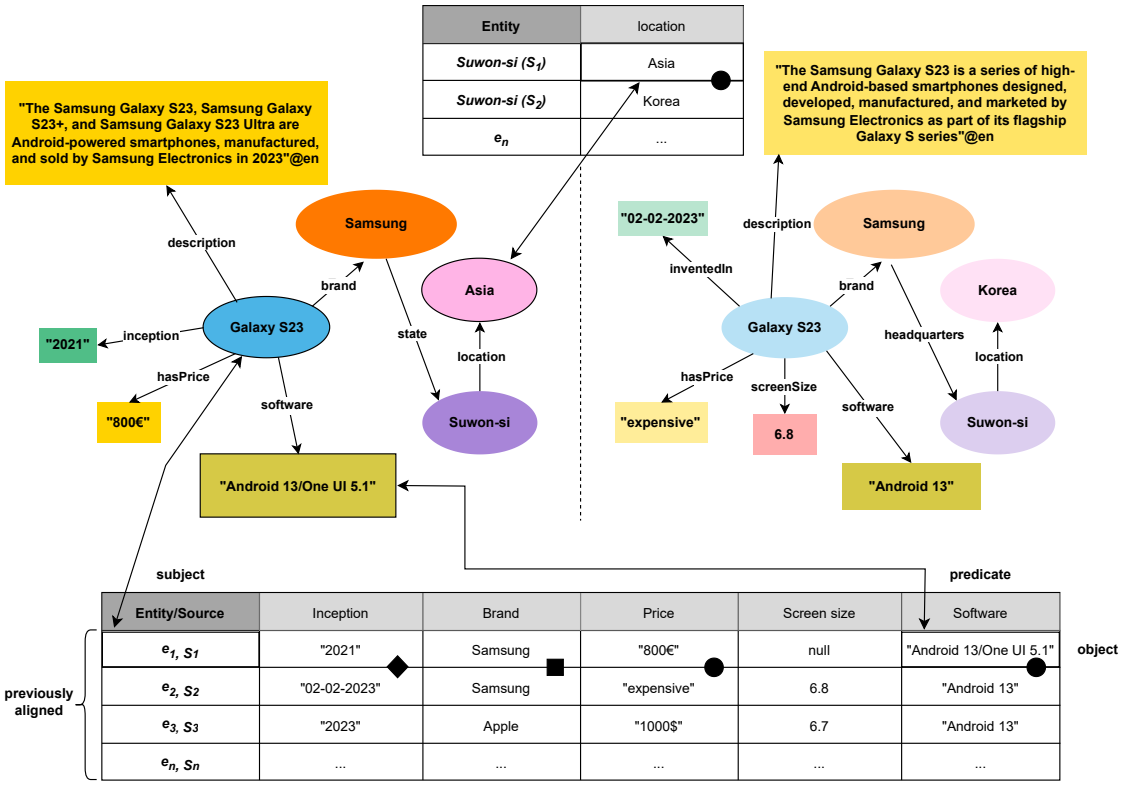

Figure 8 Extracted triples represented as graphs from two sources $\mathcal{S}_1$ and $\mathcal{S}_2$. $\mathcal{S}_1$ on the left and $\mathcal{S}_2$ on the right. With different types of conflicts: contradictory information (◆), duplicate (■), and difference in specificity (●).

- 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 involves resolving conflicts in the data to maximize these three goals. Indeed, when integrating knowledge from multiple heterogeneous sources, the quality of the information varies, and we need to determine the trustworthy information by performing a Truth Inference (TI) task. According to Rekatsinas [132], different TI strategies can be adopted. There are simple strategies that estimate the true values of the entities compared to other values provided by the sources by applying a majority vote or an average on them. Additionally, 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 metric for each class of object and for each source. The problem with simple strategies is that they do not take into account the varying quality of data sources [94], but they are often used to initialize the true values to start iterative TI methods.

For example, suppose that knowledge in the form of triples *(subject, predicate, object)* has previously been extracted about the cell 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" to refer to an element to be

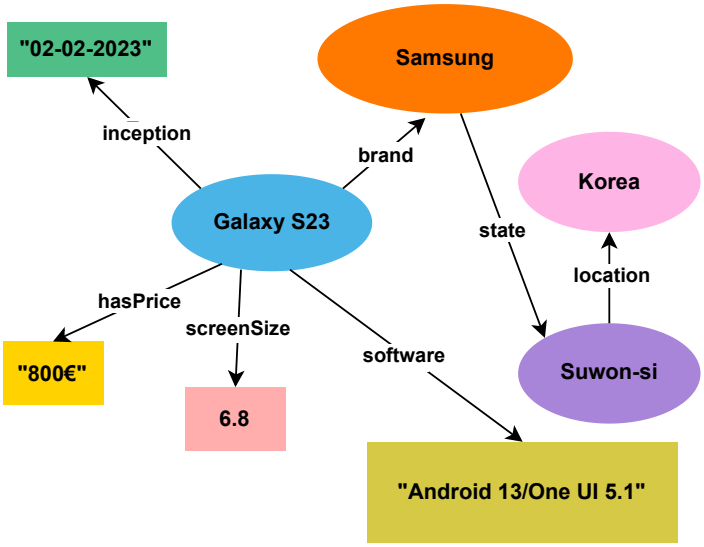

Figure 9 Resulting KG after reconciliation step (alignment and fusion).

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 in the column headers depicted in Figure 8. The data extracted from both sources are almost the same except for the relations *software* and *inception*, where the differences in specificity appear (*e.g.,* the price of the cell phone). Source $\mathcal{S}_3$ states that the brand of the cell phone is "Apple", contradicting "Samsung" provided by the first two data sources. Another example of a contradiction is the invalidity inception date of the phone provided by $\mathcal{S}_1$, which states it is 2021, whereas the correct date is 2023. We can also distinguish two levels of specificity. The first level concerns literals, such as numerical values or textual descriptions. In this case, specificity refers to the level of detail provided by the value. For example, the description of the Galaxy S23 provided by $\mathcal{S}_2$ contains more specific information than in the description provided by the first source, as illustrated in Figure 8. The second level concerns concepts. For instance, the concepts Korea and Asia may both be used to indicate the location of Suwon-si but they differ in specificity since Korea is more specific than Asia. Such differences can often be inferred from a taxonomy, as shown in Figure 10. Moreover, different levels of value representation can occur across sources, as illustrated in Figure 8. For instance, source $\mathcal{S}_1$ provides the triple ⟨Galaxy S23, hasPrice, "800€"⟩, while source $\mathcal{S}_2$ provides ⟨Galaxy S23, hasPrice, "expensive"⟩. The two values are not on the same representation scale: the first is a numerical value expressing the exact price of the cell phone, while the second is a qualitative assessment of the price. In all cases, the first and second sources provide complementary 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.

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" interchangeably.

8.2 Fusion Approaches

Fusion methods are listed in Table 3, which provides an overview and indicates whether the methods can handle certain characteristics of data or data sources such as numerical data, categorical data, data specificity, and dependencies between data sources.

SLiMFAST [132] leverages domain-specific knowledge features to improve the quality estimation of data sources. For example, if the data are 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 incorporated into the parameters of the logistic function that estimates source quality. To merge the data, they apply statistical learning to estimate source quality and then apply probabilistic inference to predict the 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 [37] incorporates the interdependence between data sources in the truth discovery process. The intuition is that a single source could provide the true value, while all the other sources could provide false values, knowing that some of them may copy from each other and therefore spread false values. Thus, if a data source provides a value different from all the others, it is not necessarily false. They define the dependency between two sources if part of their data comes directly or transitively from a common source, and it is computed using Bayesian models. Then, to discover the true value, they combine this dependency evaluation with the accuracy of the data sources, which is computed in relation to the confidences of the values and dependencies between sources.

POPACCU [41] is a refinement of the ACCU model. Unlike ACCU, it assumes that the data sources are independent and that only one value can be correct. In [41], the authors propose to select an optimal subset of data sources in order to maximize the quality of integrated data while minimizing data integration costs based on the Marginalism principle in economic theory. The model takes into account the distribution of incorrect values for an entity, based on observed values. The probability of a source providing an incorrect value depends on the popularity measure of the values. Unlike the basic model, the POPACCU variant is shown to be monotonic, *i.e.,* adding a source never reduces the accuracy of the fusion.

CRH [94] (Conflict Resolution on Heterogeneous Data) estimates the reliability of a source by using all types of data simultaneously, instead of focusing on a single type. To do this, the authors use a loss function that measures the distance between the unknown truth and the values claimed by the sources for each type of data. To initialize the source reliability score, it first applies a simple conflict resolution method, such as majority or average voting, and models the estimation of source and truth weights as an optimization problem.

MDC [97] takes into account the semantic aspect of values. To illustrate the importance of semantics, the authors provide the following example: one data source provides the true value "common cold", another source claims the value is "sinus infection", while a third claims the value is "bone fracture". Instead of examining all values at the same level, MDC calculates the semantic proximity among them. This semantic proximity calculation allows to evaluate how close a value is to the true value. The semantics of the values are captured through their vector representations, learned by following the idea that if two values share similar words, then their vectors should also be similar.

DOCS [183] is a system deployed on the Amazon Mechanical Turk, which takes into account the precision of the answers of each worker to assign tasks from specific domains to the most suitable worker. In terms of true value inference, the system uses the inherent relationships between the reliability of workers (viewed 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 for entities in the same domain as v is high, then v is likely correct; *(2)* if a source s often provides correct values for a domain d, then s has a high reliability for domain d.

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.

Model	Task		Modeling	Awareness				Datasets	Year
	TI	SQ		Numerical Data	Categorical Data	Specificity	Source dependence		
TruthFinder [176]	•	•	Probabilistic	•	•	Vagueness	•	Book authors	2007
ACCU [37]	•	•	Probabilistic	•	•	Vagueness	•	Synthetic	2009
LFC [131]	•	•	Probabilistic	•	•	○	○	Digital mammography and Breast MRI	2010
GTM [180]	•	•	Probabilistic	•	○	Fuzziness	○	Wikipedia edit history of city population and People biographies	2012
LTM [181]	•	•	Probabilistic	•	•	Completeness	○	Book author, Movie director, and Synthetic	2012
POPACCU [41]	•	•	Probabilistic	•	•	○	○	Books (from AbeBooks.com) and Flight	2012
LCA [126]	•	•	Probabilistic	•	•	○	○	Books, Population, Stocks, and FantasySCOTUS	2013
CRH [94]	•	•	Optimization problem	•	•	Vagueness	○	Weather Forecast, Stocks, and Flight	2014
CATD [93]	•	•	Optimization	•	•	○	○	City Population, Biography, and Indoor Floorplan	2014
KBT [38]	•	•	Probabilistic	•	•	○	○	Triples collected by Knowledge Vault Synthetic	2015
FaitCrowd [103]	•	•	Probabilistic	•	•	○	○	SFV and Dataset from crowdsourcing platform	2015
DOCS[183]	•	•	Probabilistic	•	•	○	○	ItemCompare, 4-Domain, Yahoo QA, and SFV	2016
ASUMS [6]	•	•	Belief functions	○	•	Fuzziness	○	Synthetic and People biographies	2016
KDEm [157]	•	•	Optimization	•	○	Completeness	○	Synthetic and Population	2016
SLiMFAST [132]	•	•	Probabilistic	•	•	○	•	Stocks, Demonstrations, Crowds, and Genomics	2017
MDC [97]	•	•	Representation Learning Optimization problem	○	•	○	○	Dataset created from baobaozhido (crowdsourcing platform)	2017
HYBRID [92]	•	•	Probabilistic	•	•	Completeness	○	Book and Synthetic	2017
TDH [81]	•	•	Probabilistic	•	•	Fuzziness	•	BirthPlaces and Heritages	2019
Record Fusion [64]	•	○	Softmax Classifier	•	•	○	○	Flight, Stock (1 & 2), Weather, and Address	2020
OKELE [17]	•	•	Probabilistic	•	•	Completeness	○	Synthetic	2020
TKGC [72]	•	•	Representation Learning Probabilistic	•	•	Completeness	○	Dataset built from [17] Subgraph of Freebase as prior knowledge (KG)	2022

TruthFinder [176] 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 verified facts. Thus, an interdependence between facts and sources emerges. 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

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

The logarithm is then applied to simplify computation and obtain the final scores. Although simple, these definitions account for the influence of facts where values with varying levels of vagueness (*e.g.,* "Joe Biden", "J. Biden", and "Biden") can increase the confidence of a fact. They also take into account the dependence between data sources through a dampening factor acting on the computation of the confidence score.

While most existing methods consider more generic values of a correct one as false, **TDH** [81] leverages the hierarchical structure of knowledge (*i.e.,* one aspect of specificity) to fuse data extracted from different sources. The idea behind this 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 modeling should not negatively affect the assessment of source reliability. Instead of classifying the value as *correct* or *incorrect*, they introduce three classifications: *exactly correct*, *hierarchically correct*, and *incorrect*. Therefore, TDH models each source not only by its reliability but also by its individual tendency to generalize or be specific, which are jointly learned in the TI process.

In the same way, **ASUMS** [6] adapts existing truth discovery models by recognizing that not all values are necessarily conflicting and identifies a partial order among 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 at different levels of specificity. Consequently, all facts more generic than a given fact are considered true, while conflicting facts are those at the same hierarchical level but with different values.

LFC [131] also measures the reliability of each source (*e.g.,* annotators in the paper) by their specificity and sensitivity relative to the unknown gold standard dataset, assigning higher weights to the best-performing sources. Estimations are iteratively performed using the EM algorithm, where the missing data is the true value initialized by applying a majority voting. The specificity and sensitivity are then estimated iteratively until convergence is reached.

LCA [126] includes four models with different sophistication: SimpleLCA, GuessLCA, MistakeLCA, and LieLCA. LCA is a probabilistic model in which the true value is represented as a multinomial latent variable. To infer the truth, the EM algorithm computes the trustworthiness of each source relative to its claims, after which the true value is determined based on the trustworthiness of the sources.

KBT [38] focuses on assessing the quality of Web sources from which facts are both extracted and 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 source reliability, distinguishing between errors arising from the facts themselves and those resulting from their extractions. However, they do not consider specificity or the possibility that multiple correct values may coexist for a single data item.

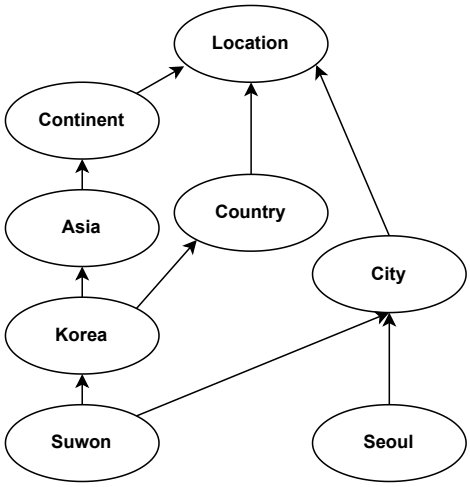

Figure 10 Illustration of a partial ordering.

While other approaches are focused on categorical data, **GTM** [180] (Gaussian Truth Model) tackles the truth finding task on numerical data. This model focuses on the relative position of numerical claim values v_c in terms of distance to find the truth. To embed the notion of distance in their model, the authors treat 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 the quality correlates with the proximity 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 mitigate the influence of outliers on the maximum likelihood estimate (MLE). Source quality is derived from a prior inverse Gamma distribution, while the truth for an entity is inferred from a prior Gaussian distribution. Finally, they use the EM algorithm to compute both the truth and source quality.

In the same manner as GTM, **LTM** [181] incorporates prior knowledge about sources or truth into the truth finding process and introduces the notion of two-sided source quality. It simultaneously infers the quality of the source and the truth, as both mutually influence each other. To assess the quality of data sources, the authors model 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 its specificity, which corresponds to the "false positive rate". These are two independent measures. Both sensitivity and specificity are modeled using a Beta distribution. For specificity, the parameters are "the prior false positive count" and "the prior true negative count", while for sensitivity they are "the prior true positive count" and "the prior false negative count". The prior probability of truth is also modeled using a Beta distribution with "the prior true count" and "the prior false count" as parameters for each distinct (entity, value) pair. The truth value is modeled using a Bernoulli distribution with parameter θ, representing the prior probability of the value being true. Finally, the truth and the quality of sources are inferred using Collapsed Gibbs Sampling.

Record Fusion [64] merges knowledge by leveraging integrity constraints, quantitative statistics, and provenance information when available. To find the true value of each table cell, it employs one classifier per column (attribute) in the dataset. These softmax classifiers can be

modular, for example a logistic regression, a decision tree, a neural network, and others. Three representation models are explored to create the feature vector, which will then be provided to the classifier. The first representation concerns the role of a cell at the column level, with three proposed strategies. The second representation 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.

In contrast to many fusion truth inferring approaches, **FaitCrowd** [103] evaluates the quality of data sources over several degrees, with one quality degree for each knowledge topic for a crowdsourcing case. FaitCrowd models the expertise of each source for each topic using a Gaussian distribution. It then models the true value provided by a source for a specific question on a given topic as a logistic function depending on the contribution ratio of the source on the topic, its expertise, and a bias term. To estimate the parameters, the model employs the Gibbs-EM inference method, which alternates between Gibbs sampling and gradient descent.

TKGC [72] takes advantage of prior knowledge from the KG when building it and considers that the noise that affects the truth is modeled by a probability distribution specific to each data source. To estimate the difference between the true value and the value provided by a source, the authors use a difference function adapted to each data type, *i.e.,* categorical, numerical, datetime, and string data. This difference function operates on representation vectors previously learned in a fact scoring function for KG completion. The output of this function follows a Gaussian distribution $\mathcal{N}(0, (k_a\sigma_s^2)$, where k_a is a regularization factor and σ_s represents the noise of the data source. Finally, TI is conducted using a semi-supervised algorithm.

OKELE [17] models the probability of a fact being true as a latent random variable following a Beta(β_0, β_1) distribution, where β_1 represents the prior true count of the fact, and β_0 represents its prior false count. The quality of a data source is characterized by its error variance ω_s, which follows a scaled inverse chi-squared distribution Scale-inv-$\chi^2(v_s, \tau_s^2)$, where v_s denotes the number of facts provided by the source, and τ_s^2 represents the variance. The authors argue that this distribution effectively handles the effect of dataset size, particularly for long-tail entities. TI is performed by leveraging prior knowledge from existing KGs to identify whether an attribute expects a single or multiple values.

HYBRID [92] addresses the TI task for knowledge in tail verticals, *i.e.,* less popular domains that are under-represented. The approach explores data fusion under two assumptions: single-truth and multi-truth. Before applying data fusion, they collect "evidences" on entities by checking whether a source contains the subject and object of the original triple. To do this, they use three types of sources: KBs (Freebase and Knowledge Vault), the Web, and query logs. Provenance information such as the URL where the system found the evidence and the pattern, is retained. Once the evidence retrieval is complete, HYBRID leverages the number of truths for each type of data item as a prior probability (for example, a cell phone has only one year of creation, or between two and eight buttons could be considered for a cell phone). 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 that when a source provides a value, a truth exists and *recall*, *i.e.,* when a truth exists, the source provides a value.

CATD [93] addresses the problem of data fusion in a context where sources provide only a few claims, making it difficult to estimate their quality. The quality of sources is modeled by a Gaussian distribution, where the mean represents the bias of the source, *i.e.,* its tendency to provide false information intentionally, and the variance represents the degree of reliability of the source. To cope with the problem of limited data from a source, they use a confidence interval of the variance to represent its reliability. Finally, CATD applies an optimization algorithm by initializing the true values with a simple method (*e.g.,* a median of the values). It first estimates the quality of the sources based on their claimed values, then iteratively refines the true values.

KDEm [157] replaces the concept of the true value with the concept of trustworthy opinions regarding the value of an entity. This model allows multiple true values for an entity's attribute, and consequently considers the completeness aspect of knowledge specificity. KDEm leverages kernel density estimation with a Gaussian kernel and extends it by incorporating the weights of the sources to estimate the probability distributions of values for each attribute of an entity. To find the 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 using different value scales such as numeric, alphanumeric, textual, or intervals of values. If these different levels of uncertainty are used for reconciliation, they must be preserved and represented in the KG as metadata to retain a history, which could potentially help resolve future conflicts [73]. The inclusion of uncertainty in the KG also enables the selection of knowledge based on its confidence level and helps maintain the quality of the graph [162]. Several works deal with querying UKGs. For example, in [59], Hartig presents tSPARQL that extends RDF model and its query language SPARQL to handle uncertainty. In [29], the authors address the failing RDF query problem, *i.e.,* when a user obtains an empty result, which can occur when querying the graph with a high confidence threshold. To do this, they use tSPARQL [59] and propose answers derived from the 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 [110], 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 [49]. Therefore, it is important to choose the best knowledge representation when building a KG according a few criteria described in Section 9.2. For instance, several sources may provide the same data, hence, the model must be capable of including all provenance information (*i.e.,* multiple data sources). For specific applications, some information to assign to triples could be required *e.g.,* provenance information, the uncertainty from extraction algorithms, or spatial and temporal information [116, 32, 61, 20, 123]. 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 [31]. 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 key notions, namely conceptualization, explicit and formal specification, and sharing [54, 53]. The conceptualization refers to 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 [173]. [1] provides a table that summarizes the usual components that form an ontology (refer to this table for further details). OWL allows the description of 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 their ability to define rich ontologies, OWL, and more generally, DLs cannot natively represent or reason about knowledge uncertainty, as they rely on crisp logic, *i.e.,* a statement is either true or false unlike fuzzy logic [30]. However, in [27], the authors argue that the inability to handle 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 and DLs [109]. These uncertainty theories include the probability theory, for example, with Bayesian network, fuzzy logic, belief functions with Dempster-Shafer (DS) theory, etc.

In [102], the authors describe the different logics available for managing uncertainty, including probabilistic logic, possibilistic logic, and many-valued logic. $\mathcal{BEL}$ [21] extends $\mathcal{EL}$ to handle uncertain knowledge with the following assumption: the KB contains information that is certain but depends on an uncertain context. One of the advantages of Bayesian networks is their capacity to represent probabilistic dependencies among different elements of a KB [21]. The Bayesian network models the probability distribution of the contexts. Each random variable in the network represents a characteristic associated with these contexts.

In [13], the authors propose a probabilistic extension of the classical DL $\mathcal{ALC}$ [3], called $\mathcal{BALC}$. The specificity of $\mathcal{BALC}$ is that axioms and assertions are annotated with an optional context, which indicates the condition for the assertion or axiom to hold. These contexts are associated with probabilities represented in a Bayesian network. Since contexts are optional, classical axioms or assertions can be encoded such that an $\mathcal{ALC}$ KB is also a $\mathcal{BALC}$ KB, with axioms and assertions that always hold.

Log-linear DLs integrate logics with probabilistic log-linear models and allow the incorporation of probabilistic and deterministic dependencies between description logic axioms [119, 118]. In [119], the authors present a log-linear DL based on $\mathcal{EL}^{++}$-LL (*i.e.,* $\mathcal{EL}^{++}$ [87] without nominals and concrete domains). An $\mathcal{EL}^{++}$-LL ontology is a pair $(\mathcal{C}^{\mathrm{D}}, \mathcal{C}^{\mathrm{U}})$, where $\mathcal{C}^{\mathrm{D}}$ is the set of deterministic axioms and $\mathcal{C}^{\mathrm{U}}$ is the set of axioms associated with a weight w representing the degree of confidence. These weights determine the log-linear probability distribution.

f-OWL [143] extends OWL DL with fuzzy set theory by adding degrees to OWL facts to represent vague knowledge. The semantics of f-OWL are provided by a fuzzy interpretation, where the interpretation function maps the elements of the domain to a membership function in $[0, 1]$, representing the degree.

OntoBayes [173] integrates Bayesian networks into OWL to leverage the advantages of both. It introduces three OWL classes: *PriorProb*, *CondProb*, and *FullProbDist* of type *ProbValue* (value between 0 and 1) to manage probabilities.

PR-OWL [27] 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) as a probabilistic model includes more than one individual (N-ary relations).

BayesOWL [30] extends OWL to represent and reason with uncertainty using 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 pairwise conditionals through two defined OWL classes *PriorProb* and *CondProb*. A prior probability for a concept is defined as an instance of the class *PriorProb* with two properties: *hasVariable* and *hasProbValue*. A conditional probability is represented as an instance of class *CondProb* with the same properties as the above instance and a property *hasCondition*.

URW3-XG [88] 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. *UncertaintyNature* categorizes uncertainty as either aleatoric, *i.e.,* ontic, or epistemic.

Poss-OWL 2 [4] extends OWL 2 to represent incomplete and uncertain knowledge from a possibilistic perspective. 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. *Axiom* corresponds to the possibilistic axioms (PossTBoxAxiom and PossABoxAxiom), where each axiom is associated with a real value representing the certainty level of the axiom. The main limitation of Poss-OWL 2 is that it focuses only on the description of uncertainty at the class level.

Riali *et al.* [134] propose a probabilistic extension of fuzzy ontologies to model vague, imprecise, probabilistic knowledge as fuzzy OWL only models vagueness. Riali *et al.* also provide a comparison of different approaches for modeling uncertainty in an ontology such as PODM [66], HyProb-Ontology [113], etc.

mUnc [34] aims to unify the different uncertainty theories within a single ontology. The ontology includes the following theories: probability, Dempster-Shafer evidence theory, 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. However, alternative graph representations can circumvent this limitation. To detail these graph representations, we consider that we want to represent the uncertainty through a confidence score $s \in [0, 1]$, where 0 indicates low confidence and 1 indicates 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.

Singleton Property [116] uses a new type of property called "singleton property", which corresponds to a unique property linked to a URI between two entities. This unique property can act as a node to which additional relations can be added. For example, the singleton property

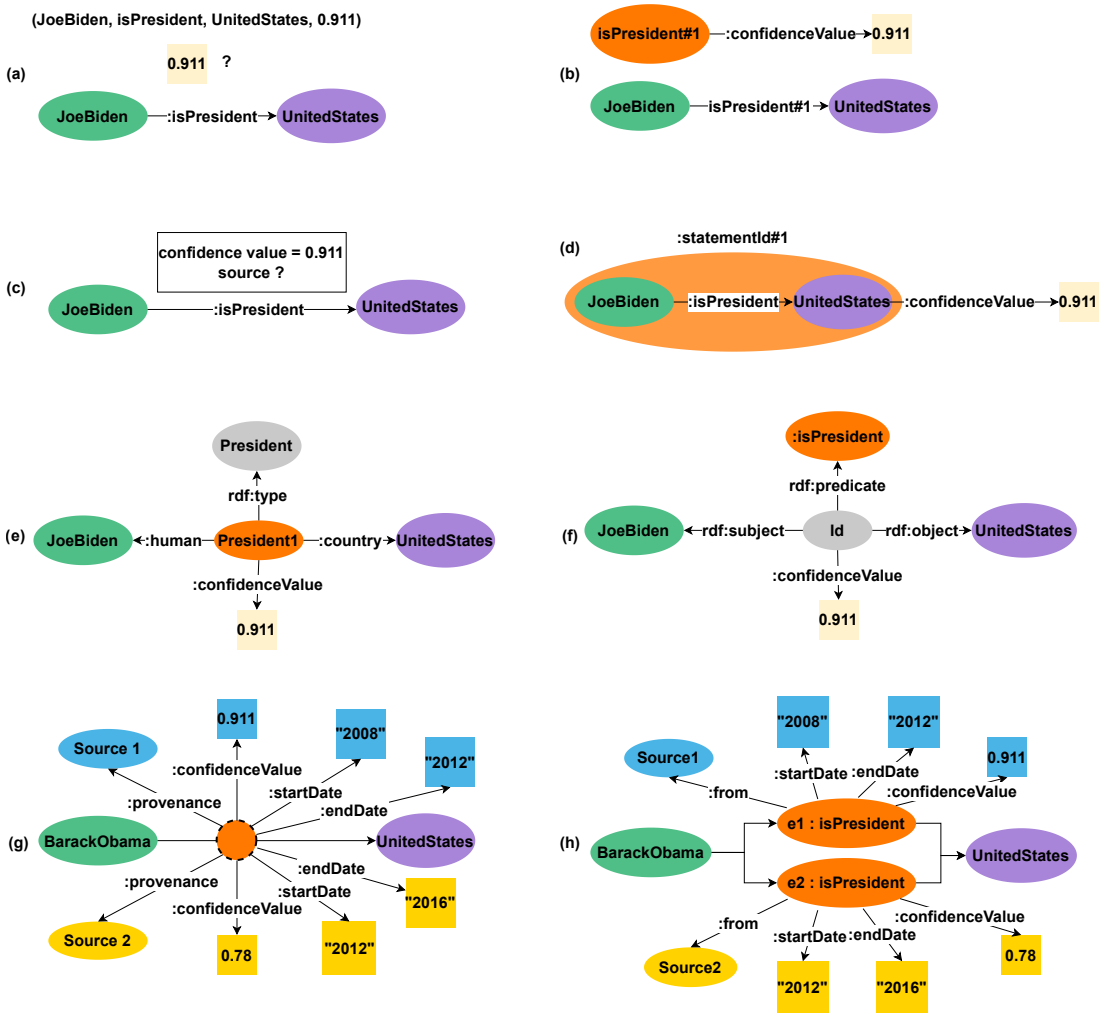

Figure 11 Illustration of different ways to embed the uncertainty (captured by a confidence score) of a triple in the KG representation: (a) RDF, (b) Singleton Property, (c) Property Graph, (d) Named Graph, (e) N-ary, (f) RDF Reification, (g) RDF-star (RDF*), (h) Multilayer Graph. ● in RDF-star illustration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

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 [135, 61].

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 [65] 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 [135].

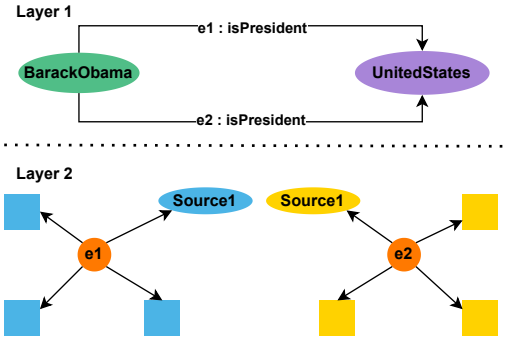

Figure 12 Illustration of the notion of "layer" where the first layer can be seen as a single triple <BarackObama, isPresident, UnitedStates> while the second layer contains different sets of information about the triple.

N-ary [104] 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 [62] involves 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 [135]. This method is simple, but its syntax is too verbose since each statement must be reified. This considerably increases the size of the KG and complicates both SPARQL querying and RDF data exchange [116, 2, 135, 60].

RDF-star [61] 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 expressively extend an existing RDF model, since the triple metadata are simply added as objects of them [80]. 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 [84]. A comparison conducted on Wikidata demonstrates 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 [80]. However, RDF-star cannot consistently represent the same metadata with different values for the same triple [60, 2]. 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 $\gamma : O \to O \times O \times O$ is a partial mapping that models directed, labeled and identified edges between objects. The layers in the multilayer graph arise 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 identifier, 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 allows the unambiguous representation of multiple provenances and varying confidence scores within a single triple.

10 Discussion and Perspectives

Throughout this survey, we have seen that there are several approaches to represent uncertainty within KGs. This is made possible by the development of ontologies that include multiple theories, enabling uncertainty to be manipulated, and by 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 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 is 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 is learned between the two embedding spaces of the UKGs to be aligned. Alternatively, confidence scores can be incorporated into the neighbor aggregation process in 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 specificity or numerical values. Knowledge specificity 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 specificity 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 specificity, 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 specificity 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. In summary, we believe that the following two points should be considered when constructing KGs to help improve their reliability:

- We believe that taking into account uncertainty at different stages (e.g. extraction, alignment, fusion) as well as provenance information could improve the traceability and quality of knowledge graphs, while enabling confidence-based KG querying.
- We advocate for the development of unified knowledge fusion models that manage and consider differences in specificity, attribute correlations, uncertainties from the previous stage of the data integration process, and heterogeneous data types in their modeling can lead to more accurate knowledge fusion for KGs.

11 Conclusion

In our current world, where knowledge may be noisy, contradictory and of different specificity, 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 specificity 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 are based on probabilistic models, and estimated both the trustworthiness of data sources and the 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 Sanjay Kumar Anand and Suresh Kumar. Uncertainty analysis in ontology-based knowledge representation. *New Gener. Comput.*, 40(1):339–376, 2022. doi:10.1007/s00354-022-00162-6.

2 Renzo Angles, Aidan Hogan, Ora Lassila, Carlos Rojas, Daniel Schwabe, Pedro A. Szekely, and Domagoj 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*, pages 11:1–11:6. ACM, 2022. doi:10.1145/3534540.3534696.

3 Franz Baader, Ian Horrocks, and Ulrike Sattler. Description logics. In Frank van Harmelen, Vladimir Lifschitz, and Bruce W. Porter, editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 135–179. Elsevier, 2008. doi:10.1016/S1574-6526(07)03003-9.

4 Safia Bal-Bourai and Aïcha 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*, volume 35 of *Procedia Computer Science*, pages 407–416. Elsevier, 2014. doi:10.1016/J.PROCS.2014.08.121.

5 Djamal Benslimane, Quan Z. Sheng, Mahmoud Barhamgi, and Henri Prade. The uncertain web: Concepts, challenges, and current solutions. *ACM Trans. Internet Techn.*, 16(1):1:1–1:6, 2016. doi:10.1145/2847252.

6 Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, and Isabelle Mougenot. How can ontologies give you clue for truth-discovery? an exploratory study. In *Proceedings of the 6th In-*

ternational Conference on Web Intelligence, Mining and Semantics, WIMS 2016, Nîmes, France, June 13-15, 2016, pages 15:1–15:12. ACM, 2016. doi:10.1145/2912845.2912848.

7 Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - A crystallization point for the web of data. *J. Web Semant.*, 7(3):154–165, 2009. doi:10.1016/j.websem.2009.07.002.

8 Howard A. Blair and V. S. Subrahmanian. Paraconsistent logic programming. *Theor. Comput. Sci.*, 68(2):135–154, 1989. doi:10.1016/0304-3975(89)90126-6.

9 Jens Bleiholder and Felix Naumann. *Conflict Handling Strategies in an Integrated Information System.* Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik, 2006. doi:10.18452/2460.

10 Jens Bleiholder and Felix Naumann. Data fusion. *ACM Comput. Surv.*, 41(1), January 2009. doi:10.1145/1456650.1456651.

11 Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie 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*, pages 1247–1250. ACM, 2008. doi:10.1145/1376616.1376746.

12 Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana 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*, pages 2787–2795, 2013. URL: https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

13 Leonard Botha, Thomas Andreas Meyer, and Rafael Peñaloza. The bayesian description logic BALC. In Magdalena Ortiz and Thomas Schneider, editors, *Proceedings of the 31st International Workshop on Description Logics co-located with 16th International Conference on Principles of Knowledge Representation and Reasoning (KR 2018), Tempe, Arizona, US, October 27th - to - 29th, 2018*, volume 2211 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2211/paper-09.pdf.

14 Khaoula Boutouhami, Jiatao Zhang, Guilin Qi, and Huan 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*, volume 1157 of *Communications in Computer and Information Science*, pages 129–136. Springer, 2019. doi:10.1007/978-981-15-3412-6_13.

15 Freddy Brasileiro, João Paulo A. Almeida, Victorio Albani de Carvalho, and Giancarlo Guizzardi. Applying a multi-level modeling theory to assess taxonomic hierarchies in wikidata. In Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 975–980. ACM, 2016. doi:10.1145/2872518.2891117.

16 Volha Bryl and Christian 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*, pages 1129–1134. ACM, 2014. doi:10.1145/2567948.2578999.

17 Ermei Cao, Difeng Wang, Jiacheng Huang, and Wei Hu. Open knowledge enrichment for long-tail entities. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 384–394. ACM / IW3C2, 2020. doi:10.1145/3366423.3380123.

18 Yixin Cao, Zhiyuan Liu, Chengjiang Li, Zhiyuan Liu, Juanzi Li, and Tat-Seng Chua. Multi-channel graph neural network for entity alignment. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1452–1461. Association for Computational Linguistics, 2019. doi:10.18653/V1/P19-1140.

19 Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, pages 1306–1313. AAAI Press, 2010. doi:10.1609/AAAI.V24I1.7519.

20 Jeremy J. Carroll, Christian Bizer, Patrick J. Hayes, and Patrick Stickler. Named graphs. *J. Web Semant.*, 3(4):247–267, 2005. doi:10.1016/J.WEBSEM.2005.09.001.

21 İsmail İlkan Ceylan and Rafael Peñaloza. The bayesian ontology language $\mathcal{BEL}$. *J. Autom. Reason.*, 58(1):67–95, 2017. doi:10.1007/S10817-016-9386-0.

22 Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 3998–4004, 2018. doi:10.24963/IJCAI.2018/556.

23 Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1511–1517. ijcai.org, 2017. doi:10.24963/IJCAI.2017/209.

24 Xuelu Chen, Michael Boratko, Muhao Chen, Shib Sankar Dasgupta, Xiang Lorraine Li, and Andrew 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*, pages 882–893. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.naacl-main.68.

25 Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. Embedding uncertain knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3363–3370. AAAI Press, 2019. doi:10.1609/aaai.v33i01.33013363.

26 Zhu-Mu Chen, Mi-Yen Yeh, and Tei-Wei Kuo. PASSLEAF: A pool-based semi-supervised learning framework for uncertain knowledge graph embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4019–4026. AAAI Press, 2021. doi:10.1609/AAAI.V35I5.16522.

27 Paulo Cesar G. da Costa, Kathryn B. Laskey, and Kenneth 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*, volume 5327 of *Lecture Notes in Computer Science*, pages 88–107. Springer, 2008. doi:10.1007/978-3-540-89765-1_6.

28 Sandra de Amo and Mônica Sakuray Pais. A paraconsistent logic programming approach for querying inconsistent databases. *Int. J. Approx. Reason.*, 46(2):366–386, 2007. doi:10.1016/J.IJAR.2006.09.009.

29 Ibrahim Dellal, Stéphane Jean, Allel Hadjali, Brice Chardin, and Mickaël Baron. Query answering over uncertain RDF knowledge bases: explain and obviate unsuccessful query results. *Knowledge and Information Systems*, 61(3):1633–1665, 2019. doi:10.1007/S10115-019-01332-7.

30 Zhongli Ding, Yun Peng, and Rong Pan. Bayesowl: Uncertainty modeling in semantic web ontologies. *Soft computing in ontologies and semantic web*, pages 3–29, 2006.

31 Renata Queiroz Dividino, Simon Schenk, Sergej Sizov, and Steffen Staab. Provenance, trust, explanations - and all that other meta knowledge. *Künstliche Intell.*, 23(2):24–30, 2009. URL: http://www.kuenstliche-intelligenz.de/fileadmin/template/main/archiv/pdf/ki2009-02_page24_web_teaser.pdf.

32 Renata Queiroz Dividino, Sergej Sizov, Steffen Staab, and Bernhard Schueler. Querying for provenance, trust, uncertainty and other meta knowledge in RDF. *J. Web Semant.*, 7(3):204–219, 2009. doi:10.1016/J.WEBSEM.2009.07.004.

33 Ahmed El Amine 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. URL: https://tel.archives-ouvertes.fr/tel-03679118.

34 Ahmed El Amine Djebri, Andrea G. B. Tettamanzi, and Fabien 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*, volume 11530 of *Lecture Notes in Computer Science*, pages 42–56. Springer, 2019. doi:10.1007/978-3-030-23182-8_4.

35 Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei 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*, pages 601–610. ACM, 2014. doi:10.1145/2623330.2623623.

36 Xin Luna Dong. Generations of knowledge graphs: The crazy ideas and the business impact. *Proc. VLDB Endow.*, 16(12):4130–4137, 2023. doi:10.14778/3611540.3611636.

37 Xin Luna Dong, Laure Berti-Équille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *Proc. VLDB Endow.*, 2(1):550–561, 2009. doi:10.14778/1687627.1687690.

38 Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949, 2015. doi:10.14778/2777598.2777603.

39 Xin Luna Dong, Alon Y. Halevy, and Cong Yu. Data integration with uncertainty. *VLDB J.*, 18(2):469–500, 2009. doi:10.1007/S00778-008-0119-9.

40 Xin Luna Dong and Felix Naumann. Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, 2(2):1654–1655, 2009. doi:10.14778/1687553.1687620.

41 Xin Luna Dong, Barna Saha, and Divesh Srivastava. Less is more: Selecting sources wisely for integration. *Proc. VLDB Endow.*, 6(2):37–48, 2012. doi:10.14778/2535568.2448938.

42 Lisa Ehrlinger and Wolfram 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*, volume 1695 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1695/paper4.pdf.

43 Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander 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*, pages 100–110. ACM, 2004. doi:10.1145/988672.988687.

44 Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer, 2007. doi:10.1007/978-3-540-49612-0.

45 Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

46 Anthony Fader, Stephen Soderland, and Oren 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*, pages 1535–1545. ACL, 2011. URL: https://aclanthology.org/D11-1142/.

47 Miao Fan, Qiang Zhou, and Thomas Fang 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*, pages 42–48. IEEE Computer Society, 2016. doi:10.1109/WI.2016.0017.

48 Nikolaos Fanourakis, Vasilis Efthymiou, Dimitris Kotzinos, and Vassilis Christophides. Knowledge graph embedding methods for entity alignment: experimental review. *Data Min. Knowl. Discov.*, 37(5):2070–2137, 2023. doi:10.1007/S10618-023-00941-9.

49 Weizhi Fei, Zihao Wang, Hang Yin, Yang Duan, Hanghang Tong, and Yangqiu Song. Soft reasoning on uncertain knowledge graphs. *arXiv preprint arXiv:2403.01508*, 2024. doi:10.48550/arXiv.2403.01508.

50 Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In Daniel Schwabe, Virgílio A. F. Almeida, Hartmut Glaser, Ricardo Baeza-Yates, and Sue B. Moon, editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 413–422. International World Wide Web Conferences Steering Committee / ACM, 2013. doi:10.1145/2488388.2488425.

51 Mikhail Galkin, Sören Auer, and Simon 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*, pages 497–502. IEEE Computer Society, 2016. doi:10.1109/WI.2016.0083.

52 Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter F. Patel-Schneider, and Ulrike Sattler. OWL 2: The next step for OWL. *J. Web Semant.*, 6(4):309–322, 2008. doi:10.1016/J.WEBSEM.2008.05.001.

53 Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993. doi:10.1006/KNAC.1993.1008.

54 Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on Ontologies*, International Handbooks on Information Systems, pages 1–17. Springer, 2009. doi:10.1007/978-3-540-92673-3_0.

55 Lingbing Guo, Zequn Sun, and Wei Hu. Learning to exploit long-term relational dependencies in knowledge graphs. In *International conference on machine learning*, pages 2505–2514. PMLR, 2019. URL: http://proceedings.mlr.press/v97/guo19c.html.

56 Peter Gärdenfors and Hans Rott. *Belief Revision*, volume 4, pages 35–132. Oxford University Press, April 1995. doi:10.1093/oso/9780198537915.003.0002.

57 Shibo Hao, Bowen Tan, Kaiwen Tang, Bin Ni, Xiyan Shao, Hengzhe Zhang, Eric P. Xing, and Zhiting Hu. Bertnet: Harvesting knowledge graphs with arbitrary relations from pretrained language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5000–5015. Association for Computational Linguistics, 2023. doi:10.18653/V1/2023.FINDINGS-ACL.309.

58 Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, and Bouchra El Asri. Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, 1:100042, 2021.

59 Olaf 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*, volume 5554 of *Lecture Notes in Computer Science*, pages 5–20. Springer, 2009. doi:10.1007/978-3-642-02121-3_5.

60 Olaf Hartig. Reconciliation of rdf* and property graphs. *CoRR*, abs/1409.3288, 2014. arXiv:1409.3288.

61 Olaf 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*, volume 1912 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-1912/paper12.pdf.

62 Patrick J Hayes and Peter F Patel-Schneider. Rdf 1.1 semantics. w3c recommendation. *World Wide Web Consortium*, 2, 2014.

63 Fuzhen He, Zhixu Li, Yang Qiang, An Liu, Guanfeng Liu, Pengpeng Zhao, Lei Zhao, Min Zhang, and Zhigang 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*, pages 367–382. Springer, 2019. doi:10.1007/978-3-030-18576-3_22.

64 Alireza Heidari, George Michalopoulos, Shrinu Kushagra, Ihab F. Ilyas, and Theodoros Rekatsinas. Record fusion: A learning approach. *CoRR*, abs/2006.10208, 2020. arXiv:2006.10208.

65 Daniel Hernández, Aidan Hogan, and Markus 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*, volume 1457 of *CEUR Workshop Proceedings*, pages 32–47. CEUR-WS.org, 2015. URL: https://ceur-ws.org/Vol-1457/SSWS2015_paper3.pdf.

66 Emna Hlel, Salma Jamoussi, Mohamed Turki, and Abdelmajid Ben 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*, volume 56 of *Smart Innovation, Systems and Technologies*, pages 243–254. Springer, 2016. doi:10.1007/978-3-319-39630-9_20.

67 Marvin Hofer, Daniel Obraczka, Alieh Saeedi, Hanna Köpcke, and Erhard Rahm. Construction of knowledge graphs: Current state and challenges. *Inf.*, 15(8):509, 2024. doi:10.3390/INF015080509.

68 Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4):71:1–71:37, 2022. doi:10.1145/3447772.

69 Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Springer, 2021. doi:10.2200/S01125ED1V01Y202109DSK022.

70 Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The even more irresistible SROIQ. In Patrick Doherty, John Mylopoulos, and Christopher A. Welty, editors, *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning, Lake District of the United Kingdom, June 2-5, 2006*, pages 57–67. AAAI Press, 2006. URL: http://www.aaai.org/Library/KR/2006/kr06-009.php.

71 Jiafeng Hu, Reynold Cheng, Zhipeng Huang, Yixiang Fang, and Siqiang 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*, pages 157–166. ACM, 2017. doi:10.1145/3132847.3132885.

72 Jiacheng Huang, Yao Zhao, Wei Hu, Zhen Ning, Qijin Chen, Xiaoxia Qiu, Chengfu Huo, and Weijun 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*, pages 956–965. ACM, 2022. doi:10.1145/3485447.3511938.

73 Ihab F. Ilyas, Theodoros Rekatsinas, Vishnu Konda, Jeffrey Pound, Xiaoguang Qi, and Mohamed 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*, pages 2259–2272. ACM, 2022. doi:10.1145/3514221.3526049.

74 Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören 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*, pages 243–246. ACM, 2019. doi:10.1145/3360901.3364435.

75 Mohamad Yaser Jaradeh, Kuldeep Singh, Markus Stocker, Andreas Both, and Sören Auer. Information extraction pipelines for knowledge graphs. *Knowl. Inf. Syst.*, 65(5):1989–2016, 2023. doi:10.1007/S10115-022-01826-X.

76 Lucas Jarnac, Miguel Couceiro, and Pierre 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*, pages 934–944. ACM, 2023. doi:10.1145/3583780.3615030.

77 Lucas Jarnac and Pierre 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*, volume 3262 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3262/paper16.pdf.

78 Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Networks Learn. Syst.*, 33(2):494–514, 2022. doi:10.1109/TNNLS.2021.3070843.

79 Shangpu Jiang, Daniel Lowd, Sabin Kafle, and Dejing Dou. Ontology matching with knowledge rules. *Trans. Large Scale Data Knowl. Centered Syst.*, 28:75–95, 2016. doi:10.1007/978-3-662-53455-7_4.

80 Robert T. Kasenchak Jr., Ahren Lehnert, and Gene 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*, volume 12739 of *Lecture Notes in Computer Science*, pages 254–260. Springer, 2021. doi:10.1007/978-3-030-80418-3_38.

81 Woohwan Jung, Younghoon Kim, and Kyuseok 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*, pages 205–216. OpenProceedings.org, 2019. doi:10.5441/002/edbt.2019.19.

82 Natthawut Kertkeidkachorn, Xin Liu, and Ryutaro Ichise. Ctranse: Confidence-based translation model for uncertain knowledge graph embedding. In *Proceedings of the Annual Conference of JSAI 33rd (2019)*, pages 1K4E105–1K4E105. The Japanese Society for Artificial Intelligence, 2019.

83 Natthawut Kertkeidkachorn, Xin Liu, and Ryutaro 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*, volume 1128 of *Advances in Intelligent Systems and Computing*, pages 170–178. Springer, 2019. doi:10.1007/978-3-030-39878-1_16.

84 Robin Keskisärkkä, Eva Blomqvist, Leili Lind, and Olaf 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*, volume 12387 of *Lecture Notes in Computer Science*, pages 37–53. Springer, 2020. doi:10.1007/978-3-030-61244-3_3.

85 Thomas N. Kipf and Max 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. URL: https://openreview.net/forum?id=SJU4ayYgl.

86 Wei Kun Kong, Xin Liu, Teeradaj Racharak, Guanqun Sun, Qiang Ma, and Le-Minh Nguyen. Weight-aware tasks for evaluating knowledge graph embeddings.

87 Markus Krötzsch, Frantisek Simancik, and Ian Horrocks. A description logic primer. *CoRR*, abs/1201.4089, 2012. arXiv:1201.4089.

88 Kenneth J. Laskey and Kathryn 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*, volume 423 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008. URL: https://ceur-ws.org/Vol-423/paper10.pdf.

89 Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. *W3C recommendation*, 30, 2013.

90 Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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*, pages 2723–2732. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1274.

91 Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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*, pages 2723–2732. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1274.

92 Furong Li, Xin Luna Dong, Anno Langen, and Yang Li. Knowledge verification for long-tail verticals. *Proceedings of the VLDB Endowment*, 10(11):1370–1381, 2017. doi:10.14778/3137628.3137646.

93 Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proc. VLDB Endow.*, 8(4):425–436, 2014. doi:10.14778/2735496.2735505.

94 Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei 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*, pages 1187–1198. ACM, 2014. doi:10.1145/2588555.2610509.

95 Shengnan Li, Xin Li, Rui Ye, Mingzhong Wang, Haiping Su, and Yingzi Ou. Non-translational alignment for multi-relational networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4180–4186, 2018. doi:10.24963/IJCAI.2018/581.

96 Xiang Li and Ralph Grishman. Confidence estimation for knowledge base population. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 396–401, 2013. URL: https://aclanthology.org/R13-1051/.

97 Yaliang Li, Nan Du, Chaochun Liu, Yusheng Xie, Wei Fan, Qi Li, Jing Gao, and Huan 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*, pages 253–261. ACM, 2017. doi:10.1145/3018661.3018688.

98 Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *SIGKDD Explor.*, 17(2):1–16, 2015. doi:10.1145/2897350.2897352.

99 Xixun Lin, Hong Yang, Jia Wu, Chuan Zhou, and Bin Wang. Guiding cross-lingual entity alignment via adversarial knowledge embedding. In Jianyong Wang, Kyuseok Shim, and Xindong Wu, editors, *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 429–438. IEEE, 2019. doi:10.1109/ICDM.2019.00053.

100 Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. *J. Web Semant.*, 76:100761, 2023. doi:10.1016/J.WEBSEM.2022.100761.

101 Qi Liu, Qinghua Zhang, Fan Zhao, and Guoyin Wang. Uncertain knowledge graph embedding: an effective method combining multi-relation and multi-path. *Frontiers Comput. Sci.*, 18(3):183311, 2024. doi:10.1007/S11704-023-2427-Z.

102 Thomas Lukasiewicz and Umberto Straccia. Managing uncertainty and vagueness in description logics for the semantic web. *J. Web Semant.*, 6(4):291–308, 2008. doi:10.1016/J.WEBSEM.2008.04.001.

103 Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei 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*, pages 745–754. ACM, 2015. doi:10.1145/2783258.2783314.

104 Frank Manola, Eric Miller, and Brian McBride. Resource description framework (rdf) primer. *W3C Recommendation*, 10(5), 2004.

105 Xin Mao, Wenting Wang, Huimin Xu, Yuanbin Wu, and Man Lan. Relational reflection entity alignment. In Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1095–1104. ACM, 2020. doi:10.1145/3340531.3412001.

106 José-Lázaro Martínez-Rodríguez, Ivan López-Arévalo, and Ana B. Ríos-Alvarado. Openie-based approach for knowledge graph construction from text. *Expert Syst. Appl.*, 113:339–355, 2018. doi:10.1016/J.ESWA.2018.07.017.

107 João P. Martins and Stuart C. Shapiro. A model for belief revision. *Artif. Intell.*, 35(1):25–79, 1988. doi:10.1016/0004-3702(88)90031-8.

108 Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren 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*, pages 523–534. ACL, 2012. URL: https://aclanthology.org/D12-1048/.

109 Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.

110 Timm Meiser, Maximilian Dylla, and Martin 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*, pages 2557–2560. ACM, 2011. doi:10.1145/2063576.2064018.

111 Pablo N. Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*, pages 116–123. ACM, 2012. doi:10.1145/2320765.2320803.

112 Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha P. Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matt Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. Never-ending learning. *Commun. ACM*, 61(5):103–115, 2018. doi:10.1145/3191513.

113 Abdul-Wahid Mohammed, Yang Xu, and Ming Liu. Knowledge-oriented semantics modelling towards uncertainty reasoning. *SpringerPlus*, 5:1–27, 2016.

114 Tapas Nayak, Navonil Majumder, Pawan Goyal, and Soujanya Poria. Deep neural approaches to relation triplets extraction: a comprehensive survey. *Cogn. Comput.*, 13(5):1215–1232, 2021. doi:10.1007/S12559-021-09917-7.

115 Hoang Long Nguyen, Dang-Thinh Vu, and Jason J. Jung. Knowledge graph fusion for smart systems: A survey. *Inf. Fusion*, 61:56–70, 2020. doi:10.1016/j.inffus.2020.03.014.

116 Vinh Nguyen, Olivier Bodenreider, and Amit 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*, pages 759–770. ACM, 2014. doi:10.1145/2566486.2567973.

117 Chien-Chun Ni, Kin Sum Liu, and Nicolas Torzec. Layered graph embedding for entity recommendation using wikipedia in the yahoo! knowledge graph. In Amal El Fallah Seghrouchni, Gita Sukthankar, Tie-Yan Liu, and Maarten van Steen, editors, *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 811–818. ACM / IW3C2, 2020. doi:10.1145/3366424.3383570.

118 Mathias Niepert. Reasoning under uncertainty with log-linear description logics. In Fernando Bobillo, Rommel N. Carvalho, Paulo Cesar G. da Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Thomas Lukasiewicz, Trevor Martin, and Matthias Nickles, editors, *Proceedings of the 7th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2011), Bonn, Germany, October 23, 2011*, volume 778 of *CEUR Workshop Proceedings*, pages 105–108. CEUR-WS.org, 2011. URL: https://ceur-ws.org/Vol-778/pospaper2.pdf.

119 Mathias Niepert, Jan Noessner, and Heiner Stuckenschmidt. Log-linear description logics. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2153–2158. IJCAI/AAAI, 2011. doi:10.5591/978-1-57735-516-8/IJCAI11-359.

120 Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried 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*, pages 3866–3878. Association for Computational Linguistics, 2018. URL: https://aclanthology.org/C18-1326/.

121 Andriy Nikolov, Victoria S. Uren, and Enrico 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, pages 185–186. ACM, 2007. doi:10.1145/1298406.1298446.

122 Natasha F. Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: Lessons and challenges. *ACM Queue*, 17(2):20, 2019. doi:10.1145/3329781.3332266.

123 Fabrizio Orlandi, Damien Graux, and Declan 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*, pages 233–240. IEEE, 2021. doi:10.1109/ICSC50631.2021.00049.

124 Sumit Pai and Luca 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*, pages 2869–2875. ijcai.org, 2021. doi:10.24963/ijcai.2021/395.

125 Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. Large language models and knowledge graphs: Opportunities and challenges. *TGDK*, 1(1):2:1–2:38, 2023. doi:10.4230/TGDK.1.1.2.

126 Jeff Pasternack and Dan Roth. Latent credibility analysis. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 1009–1020. International World Wide Web Conferences Steering Committee / ACM, 2013. doi:10.1145/2488388.2488476.

127 Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508, 2017. doi:10.3233/SW-160218.

128 Shichao Pei, Lu Yu, and Xiangliang Zhang. Improving cross-lingual entity alignment via optimal transport. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. International Joint Conferences on Artificial Intelligence, 2019. doi:10.24963/IJCAI.2019/448.

129 Zhiyuan Qi, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Yuejia Xiang, Ningyu Zhang, and Yefeng Zheng. Unsupervised knowledge graph alignment by probabilistic reasoning and semantic embedding. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2019–2025. ijcai.org, 2021. doi:10.24963/IJCAI.2021/278.

130 Zhiyuan Qi, Ziheng Zhang, Jiaoyan Chen, Xi Chen, and Yefeng Zheng. Prasemap: A probabilistic reasoning and semantic embedding based knowledge graph alignment system. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 4779–4783. ACM, 2021. doi:10.1145/3459637.3481972.

131 Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322, 2010. doi:10.5555/1756006.1859894.

132 Theodoros Rekatsinas, Manas Joglekar, Hector Garcia-Molina, Aditya G. Parameswaran, and Christopher 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*, pages 1399–1414. ACM, 2017. doi:10.1145/3035918.3035951.

133 Dave Reynolds. Position paper: Uncertainty reasoning for linked data. In *Workshop*, volume 14, 2014.

134 Ishak Riali, Messaouda Fareh, and Hafida 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*, pages 407–414. SciTePress, 2019. doi:10.5220/0007724104070414.

135 Florian Rupp, Benjamin Schnabel, and Kai 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*, volume 1686 of *Communications in Computer and Information Science*, pages 246–262. Springer, 2022. doi:10.1007/978-3-031-21422-6_18.

136 Anish Das Sarma, Xin Dong, and Alon Y. Halevy. Bootstrapping pay-as-you-go data integration systems. In Jason Tsong-Li Wang, editor, *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 861–874. ACM, 2008. doi:10.1145/1376616.1376702.

137 Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max 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*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer, 2018. doi:10.1007/978-3-319-93417-4_38.

138 August Th Schreiber, Guus Schreiber, Hans Akkermans, Anjo Anjewierden, Nigel Shadbolt, Robert de Hoog, Walter Van de Velde, and Bob Wielinga. *Knowledge engineering and management: the CommonKADS methodology*. MIT press, 2000.

139 Juan Sequeda and Ora Lassila. *Designing and Building Enterprise Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers, 2021. doi:10.2200/S01105ED1V01Y202105DSK020.

140 Kartik Shenoy, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro A. Szekely. A study of

the quality of Wikidata. *Journal of Web Semantics*, 72:100679, 2022. doi:10.1016/J.WEBSEM.2021.100679.

141 Xiaofei Shi and Yanghua Xiao. Modeling multi-mapping relations for precise cross-lingual entity alignment. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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*, pages 813–822. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1075.

142 Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975, 2016. arXiv:1612.03975.

143 Giorgos Stoilos, Giorgos B. Stamou, Vassilis Tzouvaras, Jeff Z. Pan, and Ian 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*, volume 188 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2005. URL: https://ceur-ws.org/Vol-188/sub16.pdf.

144 Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. PARIS: probabilistic alignment of relations, instances, and schema. *Proc. VLDB Endow.*, 5(3):157–168, 2011. doi:10.14778/2078331.2078332.

145 Fabian M. Suchanek, Gjergji Kasneci, and Gerhard 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*, pages 697–706. ACM, 2007. doi:10.1145/1242572.1242667.

146 Zequn Sun, Wei Hu, and Chengkai 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*, pages 628–644. Springer, 2017. doi:10.1007/978-3-319-68288-4_37.

147 Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. Bootstrapping entity alignment with knowledge graph embedding. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4396–4402, 2018. doi:10.24963/IJCAI.2018/611.

148 Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proc. VLDB Endow.*, 13(11):2326–2340, 2020. URL: http://www.vldb.org/pvldb/vol13/p2326-sun.pdf.

149 Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proc. VLDB Endow.*, 13(11):2326–2340, 2020. URL: http://www.vldb.org/pvldb/vol13/p2326-sun.pdf.

150 Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian 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. URL: https://openreview.net/forum?id=HkgEQnRqYQ.

151 Damian Szklarczyk, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T. Doncheva, Alexander Roth, Peer Bork, Lars Juhl Jensen, and Christian von Mering. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, 45(Database-Issue):D362–D368, 2017. doi:10.1093/nar/gkw937.

152 Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, and Cuiping Li. BERT-INT: A bert-based interaction model for knowledge graph alignment. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3174–3180. ijcai.org, 2020. doi:10.24963/IJCAI.2020/439.

153 Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. Entity alignment between knowledge graphs using attribute embeddings. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 297–304. AAAI Press, 2019. doi:10.1609/AAAI.V33I01.3301297.

154 Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume 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*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org, 2016. URL: http://proceedings.mlr.press/v48/trouillon16.html.

155 Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014. doi:10.1145/2629489.

156 Warren E Walker, Poul Harremoës, Jan Rotmans, Jeroen P Van Der Sluijs, Marjolein BA Van Asselt, Peter Janssen, and Martin P Krayer von Krauss. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated assessment*, 4(1):5–17, 2003.

157 Mengting Wan, Xiangyu Chen, Lance M. Kaplan, Jiawei Han, Jing Gao, and Bo Zhao. From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1885–1894. ACM, 2016. doi:10.1145/2939672.2939837.

158 Jingbin Wang, Kuan Nie, Xinyuan Chen, and Jing Lei. SUKE: embedding model for prediction in

uncertain knowledge graph. *IEEE Access*, 9:3871–3879, 2021. doi:10.1109/ACCESS.2020.3047086.

159 Jingting Wang, Tianxing Wu, and Jiatao 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*, volume 1669 of *Communications in Computer and Information Science*, pages 16–28. Springer, 2022. doi:10.1007/978-981-19-7596-7_2.

160 Xiangyu Wang, Lyuzhou Chen, Taiyu Ban, Muhammad Usman, Yifeng Guan, Shikang Liu, Tianhao Wu, and Huanhuan Chen. Knowledge graph quality control: A survey. *Fundamental Research*, 1(5):607–626, 2021. doi:10.1016/j.fmre.2021.09.003.

161 Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 349–357, 2018. doi:10.18653/V1/D18-1032.

162 Gerhard Weikum, Xin Luna Dong, Simon Razniewski, and Fabian M. Suchanek. Machine knowledge: Creation and curation of comprehensive knowledge bases. *Found. Trends Databases*, 10(2-4):108–490, 2021. doi:10.1561/1900000064.

163 Michael L. Wick, Sameer Singh, Ari Kobren, and Andrew 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*, pages 13–18. ACM, 2013. doi:10.1145/2509558.2509561.

164 Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili 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*, pages 481–492. ACM, 2012. doi:10.1145/2213836.2213891.

165 Xindong Wu, Jia Wu, Xiaoyi Fu, Jiachen Li, Peng Zhou, and Xu 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*, pages 1540–1545. IEEE, 2019. doi:10.1109/ICDM.2019.00204.

166 Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. Relation-aware entity alignment for heterogeneous knowledge graphs. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5278–5284, 2019. doi:10.24963/IJCAI.2019/733.

167 Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. Jointly learning entity and relation representations for entity alignment. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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*, pages 240–249. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1023.

168 Ruobing Xie, Zhiyuan Liu, Fen Lin, and Leyu Lin. Does william shakespeare REALLY write hamlet? knowledge representation learning with confidence. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4954–4961. AAAI Press, 2018. doi:10.1609/AAAI.V32I1.11924.

169 Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. Cross-lingual knowledge graph alignment via graph matching neural network. *arXiv preprint arXiv:1905.11605*, 2019. doi:10.18653/V1/P19-1304.

170 Bingcong Xue and Lei Zou. Knowledge graph quality management: A comprehensive survey. *IEEE Trans. Knowl. Data Eng.*, 35(5):4969–4988, 2023. doi:10.1109/TKDE.2022.3150080.

171 Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li 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. URL: http://arxiv.org/abs/1412.6575.

172 Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. Aligning cross-lingual entities with multi-aspect information. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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*, pages 4430–4440. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1451.

173 Yi Yang and Jacques 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*, pages 457–463. IEEE Computer Society, 2005. doi:10.1109/CIMCA.2005.1631307.

174 Alexander Yates, Michele Banko, Matthew Broadhead, Michael J. Cafarella, Oren Etzioni, and Stephen 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*, pages 25–26. The Association for Computational Linguistics, 2007. URL: https://aclanthology.org/N07-4013/.

175 Rui Ye, Xin Li, Yujie Fang, Hongyu Zang, and Mingzhong Wang. A vectorized relational graph convolutional network for multi-relational network alignment. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao,*

China, August 10-16, 2019, pages 4135–4141, 2019. doi:10.24963/IJCAI.2019/574.

176 Xiaoxin Yin, Jiawei Han, and Philip 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*, pages 1048–1052. ACM, 2007. doi:10.1145/1281192.1281309.

177 Lotfi A. Zadeh. Toward a generalized theory of uncertainty (gtu)–an outline. *Inf. Sci.*, 172(1-2):1–40, 2005. doi:10.1016/J.INS.2005.01.017.

178 Jiatao Zhang, Tianxing Wu, and Guilin 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*, volume 12681 of *Lecture Notes in Computer Science*, pages 256–271. Springer, 2021. doi:10.1007/978-3-030-73194-6_18.

179 Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. Multi-view knowledge graph embedding for entity alignment. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5429–5435, 2019. doi:10.24963/IJCAI.2019/754.

180 Bo Zhao and Jiawei Han. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*, 1817, 2012.

181 Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.*, 5(6):550–561, 2012. doi:10.14778/2168651.2168656.

182 Yu Zhao, Jiayue Hou, Zongjian Yu, Yun Zhang, and Qing Li. Confidence-aware embedding for knowledge graph entity typing. *Complex.*, 2021:3473849:1–3473849:8, 2021. doi:10.1155/2021/3473849.

183 Yudian Zheng, Guoliang Li, and Reynold Cheng. DOCS: domain-aware crowdsourcing system. *Proc. VLDB Endow.*, 10(4):361–372, 2016. doi:10.14778/3025111.3025118.

184 Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. A comprehensive survey on automatic knowledge graph construction. *ACM Comput. Surv.*, 56(4):94:1–94:62, 2024. doi:10.48550/arXiv.2302.05019.

185 Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Iterative entity alignment via joint knowledge embeddings. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4258–4264. ijcai.org, 2017. doi:10.24963/IJCAI.2017/595.

186 Qiannan Zhu, Xiaofei Zhou, Jia Wu, Jianlong Tan, and Li Guo. Neighborhood-aware attentional representation for multilingual knowledge graphs. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1943–1949, 2019. doi:10.24963/IJCAI.2019/269.

187 Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *arXiv preprint arXiv:2305.13168*, 2023. doi:10.48550/arXiv.2305.13168.