# Open Knowledge Base Canonicalization: Techniques and Challenges

Yang Yang, Edward Curry

*Insight SFI Research Centre for Data Analytics, University of Galway*

**Abstract**

Curated knowledge bases (CKBs) play a fundamental role in both academia and industry. They require significant human involvement to pre-define the ontology and cannot quickly adapt to new domains and new data. To solve this problem, open information extraction (OIE) methods are leveraged to automatically extract structure in the form of non-canonicalized triples <*noun phrase, relation phrase, noun phrase*> from unstructured text. OIE can be used to create large open knowledge bases (OKBs). However, noun phrases and relation phrases in such OKBs are not canonicalized, which results in scattered and redundant facts. In order to disambiguate and eliminate redundancy in such OKBs, the task of OKB canonicalization is proposed to cluster synonymous noun phrases and relation phrases into the same group and assign them unique identifiers. Nevertheless, this task is challenging due to the high sparsity and limited information of OKBs. In this paper, we provide an overview and analysis of the techniques used by the main frameworks and discuss the challenges in this topic.

**Keywords**

Knowledge Base, Knowledge Base Embedding, Clustering, Open Knowledge Base Canonicalization

## 1. Introduction

**Motivation.** The exponential growth of web data has become an indispensable source of knowledge for artificial intelligence. Mining and utilizing the knowledge of web data has been one of the key tasks of artificial intelligence in the past two decades. Pragmatically constructed from web resources, Curated Knowledge Bases (CKBs) like YAGO [1], Freebase [2], DBpedia [3], and Wikidata [4] have drawn significant attention from academia and industry [5]. Due to their effectiveness in storing and representing factual knowledge, they have been successfully applied in many real knowledge-driven applications, including knowledge reasoning [6, 7, 8, 9], question answering [10, 11, 12, 13], and recommendation systems [14, 15, 16, 17]. Early in the study of this discipline, researchers were mostly focusing on ontology design and fact expansion based on crowd-sourcing strategies. Large-scale CKBs, such as YAGO [1] and Wikidata [4], usually contain millions of entities and hundreds of millions of relational facts about them, which are stored in the form of triples (i.e., <*head entity, relation, tail entity*>). For example, in the triple <*Albert Einstein, lived in, Princeton*>, *Albert Einstein* and *Princeton* are real-world entities, and *lived in* represents the relation between *Albert Einstein* and *Princeton*. CKB construction is difficult to automate and therefore suffers from two inevitable defects: 1) human supervision, the construction of CKBs usually requires significant human supervision to pre-define the

*TEXT2KG 2024: Third International Workshop on Knowledge Graph Generation from Text, May 26–30, 2024, co-located with Extended Semantic Web Conference (ESWC), Hersonissos, Greece*

yang.yang@insight-centre.org (Y. Yang); edward.curry@universityofgalway.ie (E. Curry)

0000-0001-8190-8683 (Y. Yang); 0000-0001-8236-6433 (E. Curry)

ontology; 2) weak adaptability, this construction manner makes it unable to quickly adapt to new domains and new data generated by rapidly growing web data.

In order to solve this problem, Open Information Extraction (OIE) techniques have been utilized to improve the efficiency of knowledge base[1] construction without any pre-defined ontology. Given an unstructured text corpus, OIE methods can be used to automatically extract non-canonicalized triples in the form <*noun phrase, relation phrase, noun phrase*> requiring neither a pre-defined ontology schema nor any human supervision. This makes them highly adaptable to the knowledge from rapidly growing web data. Several prominent instances include TextRunner [18], Standford OIE [19], and MinIE [20]. Furthermore, the advent of the OIE method directly promotes the development of large-scale Open Knowledge Bases (OKBs), such as ReVerb [21], OLLIE [22, 23, 24], and OPIEC [25]. Without human supervision, this automated OIE paradigm enables OKBs to easily grow to a very large scale, thus the coverage and diversity of OKBs are much higher than CKBs.

A major shortcoming of OKBs is that, unlike CKBs, Noun Phrases (NPs) and Relation Phrases (RPs) in OKBs are not canonicalized and lack unique identifiers. Intuitively, this leads to two issues in such OKBs: 1) two NPs (or RPs) that have different surface forms but refer to the same entity (or relation) will be treated differently; 2) two NPs (or RPs) with the same surface form but referring to different entities (or relations) will be considered the same. Moreover, this shortcoming results in the storage of scattered and redundant facts, which makes OKBs extremely sparse and far from being directly used in downstream tasks.

To address the crucial shortcoming, the task of OKB canonicalization [26] was proposed to convert non-canonicalized triples in OKBs to their canonicalized form. In recent years, several methods [26, 27, 28, 29, 30, 31, 32] have been proposed, by treating this problem as a clustering task. More precisely, OKB canonicalization methods cluster synonymous NPs (or RPs) into one group and then select one NP (or RP) to represent others in the group. This task helps to disambiguate, eliminate redundancy, and integrate the highly diverse knowledge in OKBs, which can benefit downstream applications.

**Focus and Contributions.** Many comprehensive surveys have provided an overview of CKBs, in particular their construction [33, 34, 35, 36] and representation [5, 37, 38, 39, 40, 41]. Other surveys summarize methods for refinement [42, 43], completeness [44] and quality management [45]. Unlike other surveys that only focus on CKBs, we present the first overview of OKB canonicalization. Moreover, we also present practical resources and discuss future challenges. Hence, our contributions include:

• This is the first overview of OKB canonicalization, with an analysis of existing models and their approaches.

• Challenges of existing technologies in the area of OKB canonicalization are indicated as directions for future works.

**Outline.** We organize our overview as follows. Section 2 gives the background for OKB canonicalization, including definitions and examples. In Section 3, the analysis of the different OKB canonicalization techniques and challenges is discussed. Finally, the overview is concluded in Section 4.

[1]Knowledge base (KB) and knowledge graph (KG) are alternative terms in this overview.

## 2. Background

In this section, we introduce some basic concepts and then define the task of OKB canonicalization.

**Definition 1 (Curated Knowledge Base).** *A Curated Knowledge Base (CKB) is defined as a finite set of triples that are generated by human effort following the pre-defined ontology.*

Entities and relations in CKBs are well canonicalized and defined with unique identifiers. For example, the entity *Albert Einstein (German-born theoretical physicist; developer of the theory of relativity)* has unique identifiers "/m/0jcx" and "Q937" in Freebase [2] and Wikidata [4] respectively.

**Definition 2 (Open Knowledge Base).** *An open knowledge base is a finite set of non-canonicalized triples that are automatically extracted from unstructured text without any pre-defined ontology via OIE methods.*

It is noted that the definition of open knowledge base in this paper is different from [33]. In [33], they classify knowledge bases into open knowledge bases and enterprise knowledge bases, based on the organization or community. Specifically, open knowledge bases in [33] are published online and their content is accessible for the public good, such as YAGO [1] and Wikidata [4]. In contrast, enterprise knowledge bases are typically internal to a company and applied for commercial use-cases [46]. However, following our definitions of CKB and OKB, both open knowledge bases and enterprise knowledge bases in [33] are under the concept of CKBs in this paper, since they all require human effort to define the ontology but are not constructed automatically from unstructured text. Following our definition, notable examples of OKBs include ReVerb [21], OPIEC [25], and Open-CyKG [47].

To be more specific, we use Figure 1 (a) as an example. If we search an NP *Albert Einstein* in a question-answering system which is built on top of the open knowledge base in Figure 1 (a), we can only obtain answers related to *Albert Einstein* but other information hid behind *Albert* is missed, because it is unknown for machines that both *Albert Einstein* and *Albert* refer to the same entity. Moreover, it can be seen that the two non-canonicalized triples <*Albert, was born in, Ulm*> and <*Albert Einstein, was born at, Ulm*> are redundant facts, only one of which needs to be stored in practice.

**Definition 3 (OKB Canonicalization).** *Given a set of non-canonicalized triples in an OKB, the goal of OKB canonicalization is to cluster synonymous NPs referring to the same entity and synonymous RPs having the same semantic meaning into a group, which converts these non-canonicalized triples to their canonicalized forms. Then, one element is selected to represent all others in the same group.*

As shown in Figure 1 (b), the OKB canonicalization task can be divided into two subtasks: NP canonicalization (shown in blue at the top) and RP canonicalization (shown in green at the bottom). For NP canonicalization, the model should recognize that *Michael Jordan*$_1$ and *Michael Jordan*$_2$ do not refer to the same entity, and assign *Michael Jordan*$_1$ and *Michael Jeffrey Jordan* to one cluster, and assign *Michael Jordan*$_2$ to another cluster alone. Additionally, in each cluster, one NP should be selected to represent others in this cluster, such as Michael Jeffrey Jordan and Michael Jordan$_2$. Similarly, for RP canonicalization, the model needs to assign *was born in* and *was born at* into one cluster and select one (was born in) as the representative.

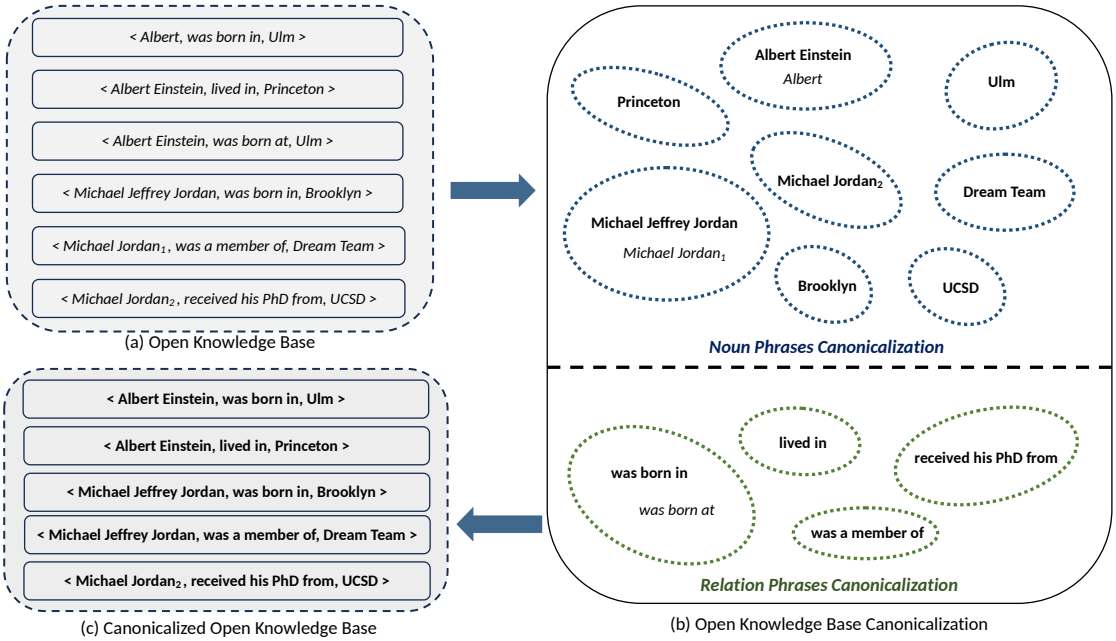


**Figure 1:** An illustration of the OKB Canonicalization task. (a) shows an open knowledge base in which neither NPs nor RPs are canonicalized, thus storing ambiguous and redundant knowledge. (b) illustrates the process of open knowledge base canonicalization, assigning NPs (or RPs) that refer to the same entity (or relation) into the same cluster, and selecting one (**bold**) to represent other elements in the cluster. (c) demonstrates a canonicalized open knowledge base without ambiguity and redundancy.

# 3. OKB Canonicalization Frameworks

In this study, the investigated OKB canonicalization models are divided into the following different categories based on the techniques utilized: (i) Side Information [48] [2], (ii) Pre-trained Word Embedding, (iii) Knowledge Base Embedding, (iv) Context Information, and (v) Supervised Manner. This section consists of an analysis of the techniques in each framework, with a discussion of their challenges. Table 1 summarizes OKB canonicalization models according to our proposed taxonomy.

## 3.1. Side Information

NPs and RPs in OKBs often have relevant side information in the documents from which the non-canonicalized triples were extracted or in other related CKBs, thus we need to use third-party tools to mine and utilize it. Furthermore, side information can be divided into two types: NP side information and RP side information.

### 3.1.1. NP Side Information

So far, previous studies have mined 14 tools to generate NP's side information. Due to space reasons, we focus here on the 3 most commonly used ones.

• **IDF Token Overlap**. Proposed by [26], this side information is also commonly leveraged by [28, 29, 49, 31, 30]. Inspired by the Inverse Document Frequency (IDF) algorithm, [26] believe that if two NPs share a word, they are more likely to be similar, but not if many other mentions

[2]It is also known as Signals in [30] and Seed Pairs in [32].

share that word. Therefore, they introduce a weighted word overlap, in which a word is given more importance if it appears in fewer mentions. It should be noted that IDF Token Overlap can not only be used for NP canonicalization but is also often used for RP canonicalization [29, 30].

• **Entity Linking**. [30] proves that OKB canonicalization and OKB linking are tightly coupled, and one task can benefit significantly from the other. Therefore, entity linking is one of the most commonly used tools for the OKB canonicalization task. For example, [26] uses FACC1 [50]; [28, 49, 31] utilize Stanford CoreNLP Entity Linker [51]; [29] leverage Wikidata Integrator [52]; and [30, 32] use Entity Popularity [53].

• **PPDB**. PPDB 2.0 [54] is a large collection of English paraphrases. All equivalent phrases will be assigned to one cluster and each cluster will randomly select a representative. Therefore, two NPs are considered equivalent if they have the same cluster representative according to the index. Similar to IDF Token Overlap, PPDB is also used for both NP canonicalization and RP canonicalization.

#### 3.1.2. RP Side Information

RP side information is obtained from relation mining and relation matching tools to improve RP canonicalization. Here, we introduce the most commonly used one, AMIE.

• **AMIE**. AMIE algorithm [55] can determine whether two RPs represent the same semantics by learning Horn rules. Previous studies [26, 28, 49, 31, 30, 32] usually take morphologically normalized triples as input to AMIE, and the output of AMIE is a set of implication rules between two RPs $p_i$ and $p_j$ (e.g. $p_i \Rightarrow p_j$ ) based on statistical rule mining. If both $p_i \Rightarrow p_j$ and $p_i \Leftarrow p_j$ satisfy the support and confidence thresholds, it means that two RPs (i.e., $p_i$ and $p_j$) have the same semantics and should be grouped into one cluster.

**Challenge 1 (Side Information).** One important challenge is to find more effective and efficient tools to boost the performance of this task. On the other hand, these third-party tools are not perfect, and the accuracy of them is usually lower than 90%. Previous methods [26, 28, 49, 31, 30, 32] usually leverage a combination of multiple side information to improve their performance. This brings a new challenge: how to evaluate the confidence of different third-party tools and use them in combination. More strategies to combine multiple tools and improve their quality should be encouraged.

### 3.2. Knowledge Base Embedding

Knowledge base embedding (KBE) [37, 38] is an increasingly popular technique that aims to represent entities and relations of knowledge bases into low-dimensional semantic spaces. To capture the relational structural information of non-canonicalized triples in an OKB, the KBE model is first exploited by CESI [28], and then it is widely used in this task. For example, HoIE [57] is used by CESI [28] and CUVA [31], and TransE [58] is utilized by CMVC [32]. Additionally, a Meta-Graph Neural Network is used in MGNN [49].

**Challenge 2 (Knowledge Base Embedding).** The main challenge in leveraging KBE on OKBs is the high sparsity of OKBs. This is because KBE can only work well on dense KBs but OKBs are usually extremely sparse. Therefore, developing tools to alleviate the sparsity of OKB and KBE methods specifically for sparse OKB could obtain improved performance. Furthermore, developing a new architecture that learns OKB embeddings while canonicalizing OKB may also

**Table 1**
Summary of OKB canonicalization models according to our proposed taxonomy. "KBE" refers to the knowledge base embedding model, "NN" refers to the Neural Network, "PWE" refers to the pre-trained word embeddings, and "CI" refers to the context information.

| Model | Side Information | | KBE | PWE | CI | Supervised Manner |
|---|---|---|---|---|---|---|
| | NP | RP | | | | |
| Galárraga et al. (CIKM2014)[26] | Attribute Overlap, String Similarity, String Identity, IDF Token Overlap, Word Overlap, Entity Overlap, Type Overlap | AMIE | - | - | - | Semi-supervised |
| CESI (WWW2018) [28] | Entity Linking, PPDB, WordNet, IDF Token Overlap, Morph Normalization | AMIE, KBP | HoIE | GloVe | - | Semi-supervised |
| SIST (ICDE2019) [29] | Jaro-Winkler Similarity, IDF Token Overlap, Entity Linking | PATTY, IDF Token Overlap | - | - | Domain Vector | Unsupervised |
| MULCE (WISE 2020) [56] | - | - | - | GloVe, BERT | - | Semi-supervised |
| MGNN (arXiv 2020) [49] | Entity Linking, PPDB, WordNet, IDF Token Overlap, Morph Normalization | AMIE, KBP | Meta-Graph NN | GloVe, BERT | - | Semi-supervised |
| CUVA (EMNLP 2021) [31] | Entity Linking, PPDB, IDF Token Overlap, Morph Normalization | AMIE, KBP | HoIE | GloVe | - | Semi-supervised |
| JOCL (SIGMOD2021) [30] | IDF token overlap, PPDB, Entity Linking | IDF token overlap, PPDB, AMIE, KBP, Relation Linking | - | FastText | - | Semi-supervised |
| CMVC (KDD2022) [32] | Entity Linking, Web Url | AMIE, Web Url | TransE | FastText | BERT | Unsupervised |

be a solution, so that high-quality embeddings could be learned during the gradual densification of OKB.

## 3.3. Pre-trained Word Embedding

Pre-trained word embeddings (PWE) have been shown to perform well in many tasks. CESI [28] first introduces GloVe embeddings [59] into this task. Specifically, before training the KBE model, all embeddings are initialized by GloVe embeddings. This is because the KBE model can only capture structural knowledge but not semantic knowledge. However, PWE can introduce semantic knowledge learned from a large corpus, thereby initializing semantically similar NPs or RPs to similar positions in the low-dimensional embedding space. Following this idea, knowledge base embeddings in CMVC [32] are initialized via FastText embeddings [60], and MULCE [56] and MGNN [49] also leveraged BERT [61] for initialization.

**Challenge 3 (Pre-trained Word Embedding).** One significant challenge for static word embeddings (such as GloVe and FastText) is the out-of-vocabulary (OOV) problem. The scale of a static word embedding dictionary is limited, but the number of words in web data is infinite (and growing all the time), thus utilizing PWE will always face the OOV problem. Besides, generating higher-quality PWE can also become a challenge. On the other hand, without fine-tuning, the performance of BERT embeddings is pretty limited but fine-tuning can be time-consuming in this task.

## 3.4. Context Information

In OKBs, non-canonicalized triples can only carry limited structural knowledge which is not enough to tackle the OKB canonicalization task, while valuable knowledge may be embedded in the source context of these non-canonicalized triples. SIST [29] first leverages knowledge from

the original source text via generating domain vectors, to cluster NPs and RPs jointly using an efficient clustering method. CMVC [32] proves that two views of knowledge (i.e., a fact view based on the non-canonicalized triples and a context view based on the non-canonicalized triple's source context) provide complementary information that is vital to the task of OKB canonicalization. Therefore, they exploit BERT [61] to convert source context into sentence embeddings and combine them with the structural embeddings of the KBE model.

**Challenge 4 (Context Information).** The first challenge is to develop more efficient models to capture the context information since only two frameworks (SIST [29] and CMVC [32]) so far exploit context information. Another challenge is how to leverage more advanced NLP models (such as generative large language models) to solve this problem.

### 3.5. Supervised Manner

For the task of OKB canonicalization, previous methods can be divided into two categories: semi-supervised methods [26, 27, 28, 30, 31] and unsupervised methods [29, 32]. Semi-supervised methods require using the validation data set to search optimal parameters. In contrast, unsupervised methods can automatically search parameters only based on the test data set without the requirement of any manually annotated label.

For example, as a clustering problem, the number of clusters (or clustering threshold) for K-means (or Hierarchical Agglomerative Clustering) is an important parameter. Previous semi-supervised methods [26, 27, 28, 30, 31] utilized the validation data set to find the optimal clustering threshold. On the contrary, SIST [29] set the clustering threshold of different data sets to the same fixed value in an unsupervised manner. To remedy this issue in a more flexible and unsupervised manner, CMVC [32] proposed the Log-Jump algorithm to predict the number of clusters, which only depends on the input embeddings of data without requiring any labels.

**Challenge 5 (Supervised Manner).** For this task, more unsupervised frameworks should be developed in the future, because we cannot always find the validation data set in real-world scenarios. In addition, for semi-supervised methods, there are differences and imbalances in probability distributions between the validation data set and the test data set. Identifying and resolving these differences and imbalances from them is also a problem. Finally, we should also encourage the development of more effective algorithms for predicting cluster numbers.

## 4. Conclusion

Open knowledge base canonicalization has become a critical topic for constructing open knowledge base from unstructured text, and this issue has not yet been fully explored. For future research, the three most important challenges can be summarized as follows: 1) How to mine and utilize more types of knowledge, such as more side information tools, knowledge from different views and modalities; 2) For frameworks that leverage knowledge base embedding models, how to alleviate the high sparsity of OKBs; 3) How to develop efficient end-to-end unsupervised frameworks.

## Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number SFI/12/RC/2289_P2.

## References


[1] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: WWW, 2007, pp. 697–706.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: SIGMOD, 2008, pp. 1247–1250.

[3] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web Journal 6 (2015) 167–195.

[4] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[5] S. Ji, S. Pan, E. Cambria, P. Marttinen, S. Y. Philip, A survey on knowledge graphs: Representation, acquisition, and applications, IEEE Transactions on Neural Networks and Learning Systems 33 (2021) 494–514.

[6] X. Chen, S. Jia, Y. Xiang, A review: Knowledge reasoning over knowledge graph, Expert Systems with Applications 141 (2020) 112948.

[7] H. Ji, P. Ke, S. Huang, F. Wei, X. Zhu, M. Huang, Language generation with multi-hop reasoning on commonsense knowledge graph, in: EMNLP, 2020, pp. 725–736.

[8] L. Liu, B. Du, Y. R. Fung, H. Ji, J. Xu, H. Tong, Kompare: a knowledge graph comparative reasoning system, in: KDD, 2021, pp. 3308–3318.

[9] Z. Li, X. Jin, W. Li, S. Guan, J. Guo, H. Shen, Y. Wang, X. Cheng, Temporal knowledge graph reasoning based on evolutional representation learning, in: SIGIR, 2021, pp. 408–417.

[10] Z. Dai, L. Li, W. Xu, Cfo: Conditional focused neural question answering with large-scale knowledge bases, in: ACL, 2016, pp. 800–810.

[11] L. Bauer, Y. Wang, M. Bansal, Commonsense for generative multi-hop question answering tasks, in: EMNLP, 2018, pp. 4220–4230.

[12] Y. Zhang, H. Dai, Z. Kozareva, A. Smola, L. Song, Variational reasoning for question answering with knowledge graph, in: AAAI, volume 32, 2018.

[13] Z. Jia, S. Pramanik, R. Saha Roy, G. Weikum, Complex temporal question answering on knowledge graphs, in: CIKM, 2021, pp. 792–802.

[14] F. Zhang, N. J. Yuan, D. Lian, X. Xie, W.-Y. Ma, Collaborative knowledge base embedding for recommender systems, in: KDD, 2016, pp. 353–362.

[15] X. Wang, X. He, Y. Cao, M. Liu, T.-S. Chua, Kgat: Knowledge graph attention network for recommendation, in: KDD, 2019, pp. 950–958.

[16] X. Wang, T. Huang, D. Wang, Y. Yuan, Z. Liu, X. He, T.-S. Chua, Learning intents behind interactions with knowledge graph for recommendation, in: WWW, 2021, pp. 878–887.

[17] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, Q. He, A survey on knowledge graph-based recommender systems, IEEE Transactions on Knowledge and Data Engineering 34 (2020) 3549–3568.

[18] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, Open information extraction from the web, in: IJCAI, 2007, pp. 2670–2676.

[19] G. Angeli, M. J. Johnson Premkumar, C. D. Manning, Leveraging linguistic structure for open domain information extraction, in: ACL, 2015, pp. 344–354.

[20] K. Gashteovski, R. Gemulla, L. Del Corro, Minie: minimizing facts in open information

extraction, in: EMNLP, 2017, pp. 2620–2630.
[21] A. Fader, S. Soderland, O. Etzioni, Identifying relations for open information extraction, in: EMNLP, 2011, pp. 1535–1545.
[22] J. Christensen, S. Soderland, O. Etzioni, An analysis of open information extraction based on semantic role labeling, in: K-CAP, 2011, pp. 113–120.
[23] M. Mausam, Open information extraction systems and downstream applications, in: IJCAI, 2016, pp. 4074–4077.
[24] S. Saha, H. Pal, Mausam, Bootstrapping for numerical open ie, in: ACL, 2017, pp. 317–323.
[25] K. Gashteovski, S. Wanner, S. Hertling, S. Broscheit, R. Gemulla, Opiec: an open information extraction corpus, in: AKBC, 2019.
[26] L. Galárraga, G. Heitz, K. Murphy, F. M. Suchanek, Canonicalizing open knowledge bases, in: CIKM, 2014, pp. 1679–1688.
[27] T.-H. Wu, Z. Wu, B. Kao, P. Yin, Towards practical open knowledge base canonicalization, in: CIKM, 2018, pp. 883–892.
[28] S. Vashishth, P. Jain, P. Talukdar, Cesi: canonicalizing open knowledge bases using embeddings and side information, in: WWW, 2018, pp. 1317–1327.
[29] X. Lin, L. Chen, Canonicalization of open knowledge bases with side information from the source text, in: ICDE, 2019, pp. 950–961.
[30] Y. Liu, W. Shen, Y. Wang, J. Wang, Z. Yang, X. Yuan, Joint open knowledge base canonicalization and linking, in: SIGMOD, 2021, pp. 2253–2261.
[31] S. Dash, G. Rossiello, N. Mihindukulasooriya, S. Bagchi, A. Gliozzo, Open knowledge graphs canonicalization using variational autoencoders, in: EMNLP, 2021, pp. 10379–10394.
[32] W. Shen, Y. Yang, Y. Liu, Multi-view clustering for open knowledge base canonicalization, in: KDD, 2022, pp. 1578–1588.
[33] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Computing Surveys 54 (2021) 1–37.
[34] G. Weikum, X. L. Dong, S. Razniewski, F. Suchanek, et al., Machine knowledge: Creation and curation of comprehensive knowledge bases, Foundations and Trends® in Databases 10 (2021) 108–490.
[35] X. Chen, H. Xie, Z. Li, G. Cheng, Topic analysis and development in knowledge graph research: A bibliometric review on three decades, Neurocomputing 461 (2021) 497–515.
[36] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A comprehensive survey on automatic knowledge graph construction, arXiv preprint arXiv:2302.05019 (2023).
[37] J. Cao, J. Fang, Z. Meng, S. Liang, Knowledge graph embedding: A survey from the perspective of representation spaces, arXiv preprint arXiv:2211.03536 (2022).
[38] G. A. Gesese, R. Biswas, H. Sack, A comprehensive survey of knowledge graph embeddings with literals: Techniques and applications., DL4KG@ ESWC 2377 (2019) 31–40.
[39] G. A. Gesese, R. Biswas, M. Alam, H. Sack, A survey on knowledge graph embeddings with literals: Which model links better literal-ly?, Semantic Web 12 (2021) 617–647.
[40] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, P. Merialdo, Knowledge graph embedding for link prediction: A comparative analysis, ACM Transactions on Knowledge Discovery from Data 15 (2021) 1–49.
[41] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches

and applications, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 2724–2743.
[42] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic web 8 (2017) 489–508.
[43] B. Subagdja, Z. Wang, A.-H. Tan, Machine learning for refining knowledge graphs: A survey, ACM Computing Surveys (2023).
[44] S. Razniewski, H. Arnaout, S. Ghosh, F. M. Suchanek, Completeness, recall, and negation in open-world knowledge bases: A survey, ACM Computing Surveys (2024).
[45] B. Xue, L. Zou, Knowledge graph quality management: a comprehensive survey, IEEE Transactions on Knowledge and Data Engineering (2022).
[46] I. Nonaka, H. Takeuchi, The knowledge-creating company, Harvard business review 85 (2007) 162.
[47] I. Sarhan, M. Spruit, Open-cykg: An open cyber threat intelligence knowledge graph, Knowledge-Based Systems 233 (2021) 107524.
[48] C. C. Aggarwal, Y. Zhao, P. S. Yu, On the use of side information for mining text data, IEEE Transactions on Knowledge and Data Engineering 26 (2014) 1415–1429.
[49] T. Jiang, T. Zhao, B. Qin, T. Liu, N. V. Chawla, M. Jiang, Canonicalizing open knowledge bases with multi-layered meta-graph neural network, arXiv preprint arXiv:2006.09610 (2020).
[50] E. Gabrilovich, M. Ringgaard, A. Subramanya, Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0) (2013).
[51] V. I. Spitkovsky, A. X. Chang, A cross-lingual dictionary for english wikipedia concepts, in: LREC, 2012.
[52] SuLab, Wikidataintegrator, in: https://github.com/SuLab/WikidataIntegrator, 2017.
[53] W. Shen, J. Wang, P. Luo, M. Wang, Linden: linking named entities with knowledge base via semantic knowledge, in: WWW, 2012, pp. 449–458.
[54] E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, C. Callison-Burch, Ppdb 2.0: better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification, in: ACL, 2015, pp. 425–430.
[55] L. A. Galárraga, C. Teflioudi, K. Hose, F. Suchanek, Amie: association rule mining under incomplete evidence in ontological knowledge bases, in: WWW, 2013, pp. 413–422.
[56] T.-H. Wu, B. Kao, Z. Wu, X. Feng, Q. Song, C. Chen, Mulce: Multi-level canonicalization with embeddings of open knowledge bases, in: WISE, Springer, 2020, pp. 315–327.
[57] M. Nickel, L. Rosasco, T. A. Poggio, Holographic embeddings of knowledge graphs, in: AAAI, 2016, pp. 1955–1961.
[58] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: NIPS, 2013, pp. 2787–2795.
[59] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: EMNLP, 2014, pp. 1532–1543.
[60] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, TACL 5 (2017) 135–146.
[61] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019, pp. 4171–4186.