

Entity Alignment for Knowledge Graphs in the Context of Supply Chain Risk Management

Rebeka Gadzo¹, Yushan Liu²

¹Humboldt University, Unter den Linden 6, 10117 Berlin, Germany

²Siemens AG, Otto-Hahn-Ring 6, 81739 Munich, Germany

Abstract

Supply chain risk management has become increasingly important during challenging times characterized by various economic, health, and political crises. To address risk mitigation issues, it is required to lay the groundwork for future risk prediction algorithms. This paper focuses on utilizing knowledge graphs, given their effectiveness in capturing and organizing complex supply chain information. The main objective is to investigate methodologies for aligning two separate knowledge graphs, where one represents supply chain data and the other external macroeconomic data. Expansion of the supply chain graph with pertinent macroeconomic information enables a better assessment and prediction of risks. The main outcome of this research work is to develop a framework based on a real-world scenario applicable to different use cases.

Keywords

Knowledge Graph, Entity Alignment, Deep Learning, Supply Chain Risk Management

1. Introduction

Supply chain risk management (SCRM) addresses strategies and their implementation for managing various risks along the supply chain based on their continuous assessment. Ensuring the highest possible extent of continuity of a supply chain and reducing its vulnerability are two of SCRM's primary goals [1].

Exogenous supply chain risks occur outside the ecosystems of companies and suppliers. They can, therefore, only be influenced by these ecosystems to a limited extent [2]. Recent notable examples are the pandemic-related effects of COVID-19 and the political and military conflict in the Ukraine. In addition, this risk type also includes legal risks, environmental and natural disasters, as well as market risks in individual regions. Exogenous risks could thus be identified through macroeconomic data such as economic facts, trends, and cause-effect relationships. The identification of these risks becomes more effective through the integration of macroeconomic

The accompanying GITHUB REPOSITORY can be found on the following LINK.


Third International Workshop on Linked Data-driven Resilience Research (D2R2'24) co-located with ESWC 2024, May 27th, 2024, Hersonissos, Greece

✉ gadzo.rebeka@gmail.com (R. Gadzo); yushan.liu@siemens.com (Y. Liu)

🌐 <https://github.com/RebekaGadzo> (R. Gadzo)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

knowledge into a company's supply chain data. Entity alignment between different data sources can facilitate this integration.

Knowledge graphs (KGs) have gained significant importance in recent years by enabling integration, management, and value extraction from diverse sources of data at large scale [3]. Additionally, they have been proven as a suitable data source for the purposes of entity alignment [4] and therefore adequate for the given problem statement.

This paper utilizes the three following techniques for entity aligning between two knowledge graphs: Dedupe [5], Ditto [6], and GPT-3 [7]. A case study on a Siemens supply chain management (SCM) KG is performed. The overall objective is to devise a general framework that can extend the Siemens SCM KG by linking it to an external macroeconomic KG. This research conducts an in-depth examination of supply chain and macroeconomic data, aiming to identify shared attributes that facilitate the dataset merging process, and assesses the performance of the three graph integration approaches.

In the introduction, the significance of entity alignment for mitigating supply chain risks is emphasized, setting the stage for the research theme. Section 2 provides an overview of existing literature in the field of entity alignment between KGs and the importance of KGs in the context of SCRM. It encompasses previous research efforts and their motivations. Section 3 describes three algorithms utilized for entity alignment between KGs, outlining their functionalities and methodologies. Section 4 focuses on the two primary data sources: the CoyPu KG¹ and Siemens' supply chain data, with an emphasis on suppliers from Germany, the USA, and China, and compares the schemas of both graphs. Section 5 details the experimental process, including the steps involved in entity alignment, and presents the research results. The conclusion summarizes the findings, discusses implications for SCRM, and outlines potential future research in the field of entity alignment between KGs in the context of SCRM.

2. Related Work

While research on entity alignment across KGs exists, there is a noticeable gap in studies specifically addressing entity alignment between KGs within SCM and SCRM contexts.

Liu et al. [8] utilized a KG to represent supply chain information from Siemens. By employing advanced KG completion techniques, the authors predicted missing connections within the graph and computed importance scores for suppliers through graph analysis algorithms. Brockmann et al. [9] focused on augmenting a supply chain KG by available web data using graph neural network models for predicting possible missing links in the created graph. Karam et al. [10] implemented Bayesian networks to propose their approach called Resilience of Supply Chain Analyzer (ReSCA), which should enable an early identification of bottlenecks in supply chains and a timely prediction of the consequences. Li et al. [11] presented the construction of a supply chain KG by performing data cleaning, desensitization, categorization, and storage of information obtained from various data sources related to the rail transportation industry's supply chain. CoyPu², a knowledge graph project within the resilience domain, operates as a semantic data platform for crisis management and supply chain data exploration. It enables

¹<https://coypu.org/ergebnisse/knowledge-graph>

²<https://coypu.org>

users to analyze complex data, make informed decisions, and contribute to economic resilience by integrating, structuring, and evaluating heterogeneous data from various sources [12].

Entity alignment is an active research area with a wide range of techniques proposed in the literature. Similarity-based methods [13], machine learning techniques [14], and deep learning models [6] have been applied to tackle entity alignment-related challenges. In recent years, the utilization of deep learning models in entity alignment tasks has become increasingly popular due to their capability to learn complex representations from data [15]. Convolutional neural networks (CNNs) [16], recurrent neural networks (RNNs) [17], and transformer-based models, such as BERT [18] and GPT-3 [19], have been employed for entity alignment tasks to capture intricate patterns in entity attributes. Ditto [6] is also considered a powerful and flexible framework for entity alignment that can be used in a wide range of applications, according to Barlaug and Gulla [20]. These deep learning models have demonstrated promising outcomes in achieving high alignment performance and robustness to noisy or incomplete data [21]. Further research is needed to develop more robust and scalable entity alignment methods to enable accurate and efficient data integration in various domains [22].

3. Algorithms

In this section, we will describe the three selected algorithms applied for the entity alignment purpose: Dedupe [5] as an unsupervised method, Ditto [6] for leveraging pre-trained language models, and GPT-3 [7] as a representative of large language models. This selection enhanced our research approach with each algorithm's distinctive strengths.

3.1. Dedupe

The Dedupe algorithm [5] employs a range of distinctive comparison methodologies. The primary technique involves utilizing the affine gap string metric to assess the similarity between diverse records. Additionally, Dedupe employs both the Levenshtein text distance and the term frequency-inverse document frequency (TF-IDF) tokens. To accurately determine field importance, the algorithm adopts regularized logistic regression. The field importance ranking process is enhanced through active learning, wherein manual input is used to improve the weight settings.

Blocking refers to the process of narrowing down candidate matches. In the case of Dedupe, it encompasses predicate blocking, which employs specific predicates, and index blocking, leveraging an inverted index for efficiency.

For grouping duplicate records, Dedupe relies on hierarchical clustering with centroid linkage. Random samples are compared against expected precision and recall metrics to set a similarity threshold for the linking process.

3.2. Ditto - Deep Entity Matching with Pre-Trained Language Models

Ditto [6] is a deep entity matching framework that leverages pre-trained language models to perform entity alignment of large datasets. The framework is built on top of PyTorch³ and the

³<https://pytorch.org/>

Hugging Face transformers library⁴. It supports both exact matching and fuzzy matching using a variety of pre-trained language models, such as BERT, RoBERTa, and DistilBERT.

Ditto uses a pre-trained language model to encode the input records into fixed-length vector representations. These vectors capture the semantic and contextual information. Consequently, the vectors are compared using a similarity function to compute a similarity score. The similarity function is based on various metrics, such as the cosine similarity. The similarity score is then thresholded to determine whether the input records are matches.

3.3. GPT-3 - Third-Generation Generative Pre-Trained Transformer

The *davinci* variant of the autoregressive language model GPT-3 [7] includes 175 billion parameters and demonstrates exceptional performance across various natural language processing (NLP) tasks, even competing with state-of-the-art fine-tuned systems. Notably, the model excels at tasks such as content generation, code writing, and translation, showcasing its potential for solving arbitrarily defined tasks and thereby establishing its utility in various NLP applications. While the architecture of the model remains similar to that of GPT-2 [23], the transformer layers display unique patterns, with the most advanced GPT-2 model having a comparatively low 1.5 billion parameters [24]. The GPT-3 model was primarily trained on English data (93%), with additional mixed-language training data. The quality of input prompts directly affects GPT-3 performance [25]. Notably, the GPT-3 model can provide a rather high confidence in its generated answers and a verbalized explanation for decision uncertainties [26].

4. Knowledge Graph Datasets

In this section, we will represent the two knowledge graph datasets used for the performed case study analysis.

4.1. Supply Chain Data

We use supply chain data from Siemens, which is stored in the form of a labeled property graph (LPG). Table 1 depicts some of the graph metadata, including the number of node types, relationship types, and the graph's size. This graph incorporates data pertaining to Siemens' suppliers, branches, and business scopes and represents a snapshot of Siemens's supply chain graph from July 2023. This KG additionally encompasses comprehensive product information, including products details, together with manufacturing and smelter information.

Siemens' supply chain boasts 61.234 suppliers in total with the most represented ones being located in Germany (6.83%), the USA (6.24%), and China (5.51%). Notably, the three mentioned countries account for approximately one fifth of the overall number of suppliers. The subsequent analysis will concentrate exclusively on the mentioned countries to facilitate more targeted model training.

⁴<https://huggingface.co>

4.2. Macroeconomic Data

CoyPu⁵ (Cognitive Economy Intelligence Platform for the Resilience of Economic Ecosystems) is a project that aims to create a macroeconomic model of the global economy in the form of a KG. It encompasses countries, industrial sectors, critical infrastructure, recent events, and further associated entities⁶. This macroeconomic data is stored as a Resource Description Framework (RDF) graph, and it is a product of several combined ontologies and data sources. The ontology can be obtained from the CoyPu project's website⁷. Table 1 shows some of the key facts about this KG.

Table 1
Key SCM and CoyPu Graph Parameters

Information	SCM	CoyPu
Graph Type	LPG	RDF
Nodes	65.568	140.128.633
Relationships	1.317.133	345.097.679
Node Types	11	83
Relationship Types	11	115
Datatype Properties	40	20
Companies	61.234	33.365.883
Graph Size	171 MB	110 GB

Among other relevant knowledge in this graph, we will emphasize the most important data such as disaster events and company properties. Alongside, there is information about natural disasters, political disasters, strikes, violence, attacks, and demonstrations. These disaster events mostly have a risk score assigned along with an alert score, which reflects the degree of their influence on a supply chain. Moreover, detailed company information can also be extracted from the KG, including company names, locations, branches, company size, and products. The KG also incorporates geographical and infrastructure information, including details such as continent, country, or city along with coordinates and the nearest ports, airports, and rivers.

5. Methodology

Entity alignment in the context of knowledge graphs entails the process of matching equivalent entities across different graphs to establish connections and integrate information effectively. This section covers the entity alignment approach between the two given data sources (see Section 4).

⁵<https://coypu.org>

⁶<https://docs.coypu.org>

⁷<https://schema.coypu.org/global/2.3>

5.1. Schema Comparison

The crucial information for entity alignment between KGs revolves around identifying identical real-world objects across two data sources. Between the macroeconomic and the supply chain data, there is a limited overlap. The chosen approach for aligning the two provided graphs involves consolidating their respective company entities. The suppliers associated with Siemens ought to be aligned with their corresponding counterparts in the CoyPu graph, thus incorporating macroeconomic details concerning these entities, including their geographical locations and potential risk associations. The country information in both graphs includes the same ISO code format, serving as a unique identifier for country matching. On the other side, some industry-related entities exhibit varying granularity levels despite representing identical information, rendering these features hardly comparable.

5.2. Framework for KGs Entity Alignment

For the purpose of entity alignment of the two data sources, three different algorithms have been applied and evaluated. Figure 1 shows the solution framework and each of the process steps. The extracted supplier lists are the starting point and serve as input data for the three selected algorithms. The results are analyzed to identify the best approach and to finally perform the envisioned entity alignment between the KGs.

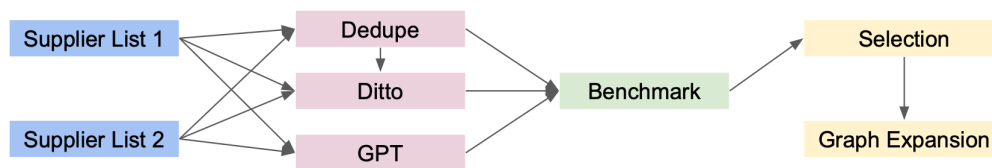


Figure 1: KGs Entity Alignment Framework

5.3. Data Preparation

For supplier matching, only the essential information describing suppliers was extracted from the initial graphs. This information, including company names, country locations, and, if available, city locations, was organized into a tabular format to facilitate the linkage process using the selected algorithms. The given data was divided into smaller units by each country to avoid memory issues, allow for country-specific data preparation and evaluation, and accelerate the speed of algorithm execution. This also resulted in a significant reduction of possible false-positive matches.

To enable the linking process, several pre-processing techniques were applied to simplify and standardize the data. Data properties were parsed, cleaned, and translated if necessary, especially to handle language or character differences. Since Chinese company names in CoyPu are in Chinese language, while SCM companies from China are already in English, the Google Translate API⁸ was employed to overcome language barriers. Notably, the company's legal form was determined based on common descriptors in company names, streamlining the alignment

⁸<https://cloud.google.com/translate/>

process by removing these descriptors from the names. Lastly, the geolocation was extracted through the external Nominatim API⁹ using city names to identify the geographical location of a majority of the listed companies. Figure 2 shows the input and output of data preparation for an exemplary Chinese company. For entity alignment purposes, anything enclosed within brackets, aside from company names, legal forms, and locations, is regarded as irregular. Geographical entities not enclosed within brackets are also identified as irregularities (see Figure 2).

Unparsed Company Name				Parsed Company Name				
Irregularity	Company Name	Legal Form	Location	Company Name	Legal Form	Longitude	Latitude	
Shanghai	Karlang Industrial	Co., Ltd.	(Taicang)	shanghai	karlang industrial	co	121.13	31.46

Figure 2: Initial and Parsed Company Name

The training process of the Ditto model requires labeled training data. The number of possible matches between the two datasets is relatively low compared to the total number of candidate pairs. Possible matches account for the number of SCM suppliers as a smaller dataset with an approximate count of 11.000 out of more than 5 billion possible candidate pairs. Manually creating viable training sets was infeasible due to the large number of possible matches. Therefore, we used the output of Dedupe to create labeled datasets for training. The proposed matches from this matching approach were additionally manually labeled.

Additionally, we have omitted the legal form from the input data of GPT-3, as previous tests have shown improved results through this approach.

5.4. Entity Alignment Implementation

The source code repository is available on Github¹⁰. It was necessary to start with the execution of the unsupervised approach Dedupe to find positive matches for data labeling. Negative matches were generated randomly and additionally labeled manually. This data has been used to provide the training set for the supervised method Ditto.

The Ditto algorithm was applied next. For each of the three countries, a corresponding model has been trained. The initial step was to generate the blocking data for matching purposes, which was done using the pre-trained model *distilbert-base-uncased-finetuned-sst-2-english* for US and Chinese companies, which were previously translated to English. In the case of German companies, *bert-based-german-cased* was used. These pre-trained models have been utilised from the Hugging Face transformers library. The blocking process was performed separately for each country.

The *gpt-3.5-turbo* API provided by OpenAI¹¹, in the following referred to as GPT-3, has also been applied for the company alignment. System messages are used to configure GPT-3 to function as a company matching service, while user messages have been utilized to send candidates pairs. The model returns a binary match value, indicating whether the two companies are a match. Alongside, the match confidence is indicated by a decimal value ranging from 0 to 1, with 1 representing full confidence in the model’s decision. The model is capable of matching

⁹<https://nominatim.org/>

¹⁰https://github.com/RebekaGadzo/graph_alignment

¹¹<https://openai.com/>

with or without specified company’s coordinates. It uses a single model for all countries, which is distinct from the other two solutions.

6. Results

This section encompasses results’ discussion based on different evaluation metrics along with the final entity alignment of the given data sets.

6.1. Results Evaluation

The test set contains 750 entries across the countries, having 231 German companies, 252 Chinese companies, and 267 US companies. It is based on a random sample of positive matches from Dedupe’s outputs, which are manually labeled for the purpose of evaluation. Negative matches were generated using random pairs from the datasets which are manually labeled.

Table 2
Results across all Countries

	True Pos.	False Neg.	True Neg.	False Pos.	Acc.%	Prec.%	Recall%	F1%
Dedupe	219	54	433	44	86.93	83.27	80.22	81.72
Ditto	260	13	361	116	82.80	69.15	95.24	80.12
GPT-3	227	46	450	27	90.27	89.37	83.15	86.15
	Pos.=273		Neg.=477		Total=750			

The models exhibited varying degrees of performance regarding false-positive and false-negative values across all three countries (see Table 4). We hypothesize that the inflated number of false-positives overall in Table 2 was due to the company name convention irregularities observed in China (see Figure 2 and Table 4). Specifically, the Ditto model matching results contained an increased number of false-positive matches, whereas the remaining two models have faced challenges with false-negative matches. This is evident from the contrasting differences in precision and recall scores of these models (see Table 2).

Table 3
Dedupe’s Matches per Country

	SCM Suppliers	Matches	Percent
Germany	4184	2341	55.95%
US	3819	452	11.84%
China	3374	224	6.64%
Total	11377	3017	26.52%

The GPT-3 model exhibited consistent performance across all countries as measured by the F1-score. Its application was found to have superior performance, but due to its high usage expenses, it is impractical for the application. As a result, the Dedupe method emerged as the

most suitable for aligning the two given graphs based on its favorable balance between recall and accuracy scores.

Table 4
Results per Country

GER	True Pos.	False Neg.	True Neg.	False Pos.	Acc.%	Prec.%	Recall%	F1%
Dedupe	71	25	131	4	87.45	94.67	73.96	83.04
Ditto	93	3	104	31	85.28	75.00	96.88	84.55
GPT-3	88	8	121	14	90.48	86.27	91.67	88.89
	Pos.=96		Neg.=135		Total=231			
USA	True Pos.	False Neg.	True Neg.	False Pos.	Acc.%	Prec.%	Recall%	F1%
Dedupe	100	16	149	2	93.26	98.04	86.21	91.74
Ditto	111	5	145	6	95.88	94.87	95.69	95.28
GPT-3	93	23	142	9	88.01	91.18	80.17	85.32
	Pos.=116		Neg.=151		Total=267			
CHN	True Pos.	False Neg.	True Neg.	False Pos.	Acc.%	Prec.%	Recall%	F1%
Dedupe	48	13	153	38	79.76	55.81	78.69	65.31
Ditto	56	5	112	79	66.67	41.48	91.80	57.14
GPT-3	46	15	187	4	92.46	92.00	75.41	82.88
	Pos.=61		Neg.=191		Total=252			

Table 3 shows the number and percentage of supplier matches found in total and per country by Dedupe, which we selected as the most suitable method. The most matches originate from Germany, while Chinese suppliers were the least likely to be matched.

6.2. Knowledge Graph Entity Alignment

The KG entity alignment includes two steps: the suppliers' property expansion and risk events addition (see Figure 3).

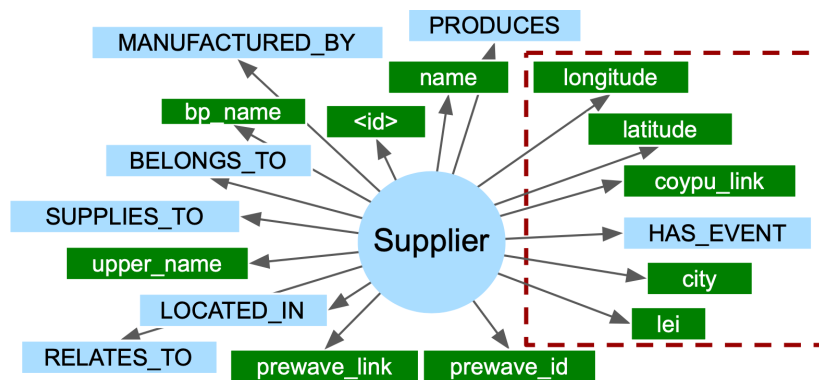


Figure 3: New Supplier's Properties

We have introduced four properties to the SCM KG: CoyPu link as ID reference of the respective company in the CoyPu KG, legal entity identifier (LEI), city name of the company's location, and latitude and longitude of the company's location.

It should be noted that although the CoyPu ontology includes many company properties, many values are still missing. These additional properties could be easily introduced to the SCM graph as the CoyPu graph completion progresses, based on the newly introduced CoyPu link. We have also established relations in the graph between events and companies that are geographically close based on leveraged coordinate information from both companies and events.

7. Conclusion

Siemens and CoyPu supplier entity alignment was conducted utilizing a set of three distinct methods. A KG entity alignment framework was developed and applied. The initial identification of matches was accomplished through the application of Dedupe, which was crucial for providing training data that could not have been feasibly generated manually.

Although GPT-3 was the most reliable method, its API cost rendered it impractical for the given problem statement. However, as large language models advance, this concern may diminish. Dedupe exhibited commendable matching capabilities with the second-best performance, demonstrating a balanced performance and practicality. Notably, the success rate of the models varied across different countries, with China posing the most problematic alignment scenario.

The outcome of the entity alignment process resulted in the Siemens supply chain graph not only containing references to the corresponding companies within the CoyPu graph but also incorporating relevant risk events associated with these companies. Future expansions may incorporate additional CoyPu graph data to enhance supply chain reliability. Our framework can be further applied to other graphs and extended by the application of risk prediction algorithms.

Acknowledgements

This work has been supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) as part of the project CoyPu under grant number 01MK21007K.

References

- [1] A. Adhitya, R. Srinivasan, I. A. Karimi, Supply chain risk identification using a HAZOP-based approach, *AIChE Journal* 55 (2009) 1447–1463. doi:10.1002/aic.11764.
- [2] S. DuHadway, S. Carnovale, B. Hazen, Understanding risk management for intentional supply chain disruptions: Risk detection, risk mitigation, and risk recovery, *Annals of Operations Research* 283 (2019) 179–198.
- [3] 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 (CSUR)* 54 (2021) 1–37.
- [4] R. Zhang, B. D. Trisedy, M. Li, Y. Jiang, J. Qi, A benchmark and comprehensive survey on knowledge graph entity alignment via representation learning, 2022. arXiv:2103.15059.

- [5] F. Gregg, D. Eder, H. Cushman, Entity resolution with machine learning: Dedupe.io's scalable foundation for data quality, 2018.
- [6] Y. Li, J. Li, Y. Suhara, A. Doan, W.-C. Tan, Deep entity matching with pre-trained language models, arXiv preprint arXiv:2004.00584 (2020).
- [7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners (2020). arXiv:2005.14165.
- [8] Y. Liu, B. He, M. Hildebrandt, M. Buchner, D. Inzko, R. Wernert, E. Weigel, D. Beyer, M. Berbalk, V. Tresp, A knowledge graph perspective on supply chain resilience, in: The 2nd International Workshop on Linked Data-Driven Resilience Research, Extended Semantic Web Conference, 2023.
- [9] N. Brockmann, E. Elson Kosasih, A. Brintrup, Supply chain link prediction on uncertain knowledge graph, ACM SIGKDD Explorations Newsletter 24 (2022) 124–130.
- [10] N. Karam, S. Matini, R. Laas, T. Hoppe, A hybrid knowledge graph and bayesian network approach for analyzing supply chain resilience, in: European Semantic Web Conference, Springer, 2023, pp. 27–31.
- [11] S. Li, Y. Zhang, M. Huang, H. Wu, W. Cai, Building and using a supply chain knowledge graph applied to the rail transit industry, in: 2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT), 2021, pp. 742–746. doi:10.1109/ACAIT53529.2021.9731237.
- [12] S. Bin, C. Stadler, N. Radtke, K. Junghanns, S. Gründer-Fahrer, M. Martin, Base platform for knowledge graphs with free software (2023).
- [13] O. Lauzanne, J.-S. Frenel, M. Baziz, M. Campone, J. Raimbourg, F. Bocquet, Optimizing the retrieval of the vital status of cancer patients for health data warehouses by using open government data in france, International Journal of Environmental Research and Public Health 19 (2022). doi:10.3390/ijerph19074272.
- [14] M. Paganelli, F. D. Buono, M. Pevarello, F. Guerra, M. Vincini, Automated machine learning for entity matching tasks, in: International Conference on Extending Database Technology, 2021.
- [15] J. Kim, K. Kim, M. Sohn, G. Park, Deep model-based security-aware entity alignment method for edge-specific knowledge graphs, Sustainability 14 (2022). doi:10.3390/su14148877.
- [16] N. T. Tam, H. T. Trung, H. Yin, T. Vinh, D. Sakong, B. Zheng, N. Q. V. Hung, Multi-order graph convolutional networks for knowledge graph alignment, in: 37th IEEE international conference on data engineering, Chania, Greece, 2021, pp. 19–22.
- [17] L. Guo, Z. Sun, W. Hu, Learning to exploit long-term relational dependencies in knowledge graphs (2019). arXiv:1905.04914.
- [18] M. Paganelli, F. D. Buono, A. Baraldi, F. Guerra, Analyzing how BERT performs entity matching, Proc. VLDB Endow. 15 (2022) 1726–1738.
- [19] J. Tang, Y. Zuo, L. Cao, S. Madden, Generic entity resolution models, in: NeurIPS 2022 First Table Representation Workshop, 2022.
- [20] N. Barlaug, J. A. Gulla, Neural networks for entity matching: A survey, ACM Transactions

on Knowledge Discovery from Data (TKDD) 15 (2021) 1–37.

- [21] K. Zeng, C. Li, L. Hou, J. Li, L. Feng, A comprehensive survey of entity alignment for knowledge graphs, *AI Open* 2 (2021) 1–13.
- [22] Q. Zhu, H. Wei, B. Sisman, D. Zheng, C. Faloutsos, X. L. Dong, J. Han, Collective knowledge graph multi-type entity alignment, in: *The Web Conference 2020*, 2020.
- [23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, *OpenAI Blog* (2019).
- [24] I. Solaiman, M. Brundage, J. Clark, A. Askill, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, J. Wang, Release strategies and the social impacts of language models (2019). [arXiv:1908.09203](https://arxiv.org/abs/1908.09203).
- [25] M. Bavarian, H. Jun, N. Tezak, J. Schulman, C. McLeavey, J. Tworek, M. Chen, Efficient training of language models to fill in the middle (2022). [arXiv:2207.14255](https://arxiv.org/abs/2207.14255).
- [26] S. Lin, J. Hilton, O. Evans, Teaching models to express their uncertainty in words (2022). [arXiv:2205.14334](https://arxiv.org/abs/2205.14334).