Research Statement

Djellel Difallah (djellel@nyu.edu)

My research addresses the challenge of developing effective methods to reason over large, incomplete, and heterogeneous knowledge bases, aiming to enhance knowledge completeness and discovery to support informed decision-making. In an era of information overload, extracting reliable insights from noisy and fragmented knowledge sources is increasingly difficult. Navigating complex knowledge ecosystems requires the development of new methods to identify information gaps, integrate diverse data sources and modalities, and interpret the reasoning of AI models when applied to these knowledge bases. My approach emphasizes creating interpretable systems that adapt to evolving domain knowledge and integrate seamlessly into human-centric workflows. Specifically, my research focuses on the following technical challenges:

- ⋆ **Knowledge Completion.** Develop scalable, automated methods to identify and fill gaps in large knowledge bases, integrating textual content with structured, graph-based data representations to enhance data completeness.
- ⋆ **Explainable Question Answering over Knowledge Graphs.** Design systems that provide explicit reasoning paths derived from knowledge graphs to answer natural language queries, ensuring the answers are interpretable and grounded in the data.
- ⋆ **Integration of Heterogeneous and Multimodal Data.** Develop models and frameworks capable of effectively combining multiple modalities of data: graph, textual, relational, visual, and temporal, to improve various tasks such as knowledge completeness and question answering.

1 Knowledge Completion

1.1 Motivation and Research Challenges

Every day, millions of people use search engines to seek information online. Many search results depend on knowledge bases, which are structured repositories of interconnected information. These knowledge bases range from curated databases maintained by institutions to collaborative platforms built and maintained through crowdsourcing. Wikipedia exemplifies the latter approach; it is the largest encyclopedia built entirely by human contributors, with hundreds of thousands of volunteers worldwide making more than 15 million edits per month.

Another prominent project is Wikidata, a collaborative knowledge graph that represents structured data as *triples*, i.e., (subject → property → object), which encode the attributes of entities, and the relationships among them. For example, (Berlin → areaSize → 891.8 km^2) and (Berlin → capitalOf → Germany), respectively. This triple-based structure organizes structured information into a graph, enabling efficient querying and supporting logical reasoning. As of 2024, Wikidata contains over 1.5 billion triples describing approximately 120 million unique entities. Moreover, Wikidata and Wikipedia are tightly interconnected by design; each Wikipedia article links directly to a unique language-agnostic Wikidata entity, bridging multilingual textual content with structured, entity-centric knowledge.

These collaborative knowledge bases illustrate key research challenges addressed by my work: (i) they contain substantial data contributed by users with diverse expertise, languages, and cultural backgrounds; (ii) their data is inherently incomplete, noisy, and nuanced; (iii) these platforms serve as authoritative sources for various downstream applications, including search engines, question-answering systems, and AI models requiring up-to-date knowledge; (iv) the pace of information creation is bound to exceed the community's capacity to monitor and maintain these projects. These characteristics make Wikipedia and Wikidata ideal testbeds for developing automated systems that identify, prioritize, and assist in knowledge completion tasks (i.e., detecting and filling information gaps), leveraging machine learning to support human curation.

1.2 Contributions

Cross-Lingual Structure and Link Completion. The quality of a text-based knowledge base depends on sourced, well-organized, and interconnected content. In Wikipedia, two organizational features facilitate information discovery: *hyperlinks*, which connect related concepts, and *article sections*, which break content into logical segments. However, despite community efforts, hyperlinks remain extremely sparse and unequally distributed, especially in low-resource languages with limited editor bases. Additionally, article structure vary significantly both within and across languages due to differing editorial norms. For instance, the English article about Berlin includes a "Demographics" section, whereas the German version uses "Bevölkerung" (population). Although these titles are not literal translations, they are used to annotate the same information. Consequently, such structural misalignment hinders cross-language knowledge transfer.

To address structural inconsistencies, we define a new task called *section title mapping* for cross-language knowledge alignment and develop a machine learning model capable of classifying corresponding section titles across languages [3]. Our model involves training a classifier on features derived from section titles and content

(e.g., word overlap, entity mentions, and embedding similarity). A global consistency solver further refines the alignments produced by the initial classifier across all language versions by enforcing constraints such as transitive matching and synonym relationships. Using these alignments, our system can identify content gaps across various language editions. For example, it can flag that the French Wikipedia article on "Climate Change" is missing information about renewable energy. The system can also be utilized to suggest article templates.

For hyperlink sparsity, we developed AddLink [6], an entity linking system that uses machine learning to automatically identify potential hyperlink placements in articles. Our approach combines a language-agnostic entity linking model that leverages data from millions of links collected across wiki-projects, and an ensemble-based machine learning model using language-agnostic features such as cross-lingual entity embeddings and statistical link distributions. The system processes the entire Wikipedia and ranks candidate links, which are then presented through an interactive interface to active contributors.

Based on such efforts, we established a generic human-in-the-loop workflow that frames Wikipedia edit actions that can be semi-automated by (i) having the system perform in the background the heavy lifting required to identify the most promising actions, while (ii) the human editors swiftly verify and validate the recommendations with additional context. This paradigm is particularly effective at lowering the entry barrier for newcomers while preserving editorial quality control.

Text-grounded Knowledge Graph Completion. Wikipedia restricts its articles to notable entities, whereas Wikidata is more permissive, allowing the storage of any entity. Consequently, most entities in Wikidata are considered *tail* entities characterized by sparse information, and thus, remain poorly connected to the broader knowledge graph. In contrast, entities covered extensively by dedicated Wikipedia articles typically exhibit richer interconnections, highlighting a clear correlation between textual coverage and structural connectivity within knowledge graphs. As the Wikidata graph expands in entity count, the potential number of connections grows quadratically, surpassing the capacity of volunteer editors to manually complete and curate the graph. Automated methods to enrich and complete the graph become essential, especially when relying on structured data alone is insufficient.

To address entity sparsity, we developed ParaGraph [14], a reverse-linking method that maps Wikidata entities to relevant Wikipedia passages. ParaGraph first generates concise summaries for each entity by combining its labels, attributes, and relationships, then identifies relevant paragraphs by embedding both entities and paragraphs into a shared semantic space. A subsequent ranking stage further refines these matches using contextual features, such as section titles and graph path statistics. By augmenting tail entities with textual context, ParaGraph enriches entity information, creating new opportunities for downstream applications.

Additionally, the *link prediction* task plays a crucial role in knowledge graph completion by directly modeling graph structural patterns. Existing neural network-based models, which rely exclusively on structural information, can effectively capture structural patterns, from simple symmetric relationships, for example, $(A \rightarrow \text{spouse} \rightarrow B)$ implying $(B \rightarrow \text{spouse} \rightarrow A)$, to more complex relational structures. However, research suggests that combining structural and textual evidence significantly enhances link prediction performance [15]. To complement entity structural information with textual evidence, we introduce WikiRAG [4], a retrieval-augmented generation (RAG) framework that grounds link prediction in textual content. Our approach uses a base knowledge graph completion model to generate initial candidate rankings, converts predictions into targeted pseudo-questions, employs dense passage retrieval to extract relevant contextual information, and applies large language model reasoning with domain consistency constraints. WikiRAG represents the first application of retrieval-augmented generation to knowledge graph completion, and has demonstrated strong performance gains on the link prediction task.

1.3 Impact

My work on knowledge completion has been published in top venues, including CIKM 2021, IEEE BigData 2023, and KDD 2025. Most significantly, the AddLink system was deployed as part of Wikipedia's production infrastructure in collaboration with the Wikimedia Foundation's Growth and Research teams. It actively supports the onboarding of newcomers into the community, ultimately increasing contribution and retention in Wikipedia's volunteer community. The research has also generated valuable community resources: ParaGraph produced a dataset for entity-to-text mapping [13] in collaboration with the University of Fribourg (Prof. Philippe Cudre-Mauroux and Natalia Ostapuk), while WikiRAG introduced an updated Wikidata benchmark addressing limitations in existing evaluation frameworks and promoting the use of RAG for link prediction. Additionally, this work has fostered interdisciplinary collaboration, including a grant from NYUAD CITIES Research Center, in collaboration with Prof. Etienne Wasmer (NYUAD Social Science), to build collaborative platforms for collecting structured data to analyze urban phenomena.

2 Explainable Question Answering over Knowledge Graphs

2.1 Motivation and Research Challenges

With the rapid growth of data generated from various sources, organizations increasingly turn to knowledge graphs to manage, query, and publish data [8]. Efforts to make government, scientific, and institutional datasets available through open data initiatives have fueled the expansion of knowledge graph data repositories, and as a result, the volume of data in this format is expected to increase significantly. These initiatives are transforming isolated data silos into interconnected knowledge ecosystems of an unprecedented scale; and these can be queried, combined, and reasoned over by anyone. In this context, knowledge graphs (KGs) have emerged as powerful representations for integrating heterogeneous datasets. Unlike traditional database systems, knowledge graphs capture fine-grained, semantically rich relationships between entities. These rich semantic relationships make it possible to perform advanced reasoning that uncovers hidden patterns, supports informed decision-making, and enables complex question answering.

However, the full potential of KGs remains largely unfulfilled due to several challenges [7], including (i) Accessibility: effectively querying knowledge graphs typically requires expertise in specialized query languages (e.g., SPARQL), leaving both domain and non-domain experts unable to directly interact with these rich data resources; (ii) Black-box reasoning: Neural methods, such as graph neural networks, often achieve strong performance but operate as black boxes, making it challenging to verify their outputs; (iii) Uninterpretable responses: While conversational AI models such as ChatGPT can assist with reasoning over structured data, they are not designed to provide a systematic explanation when facing new or proprietary datasets and hence become prone to hallucination; (iv) Data incompleteness: KGs, particularly collaborative ones such as Wikidata, are inherently incomplete and may be missing information that affects the reasoning process. My research addresses these challenges through building natural language interfaces and systems that provide explicit, verifiable reasoning.

2.2 Contributions

Reasoning with Incompleteness. To address knowledge graph incompleteness, we leverage Knowledge Graph Embeddings (KGEs), which map entities and relations into continuous vector spaces to infer missing links. KGEs scale gracefully to million-node knowledge graphs and offer low-latency responses critical for question-answering applications. However, traditional KGE methods use Euclidean spaces, which poorly capture the hierarchical relationships inherent in many knowledge graphs.

We developed HyperKGQA [12], which embeds knowledge graphs in hyperbolic space to better capture hierarchical relationships. Our system aligns natural language questions with this hyperbolic representation, enabling more accurate question answering over incomplete knowledge graphs by leveraging both existing and inferred links. In our experiments, we demonstrate that using hyperbolic embeddings for question answering significantly outperforms alternative embeddings when the target knowledge graph exhibits hierarchical properties.

Interpretability with Reasoning Paths. To support interpretability in question answering, we aim to develop systems that return along with each answer an explicit reasoning path through the graph that users can inspect to validate the answer. For example, given the question "Who wrote the book that inspired The Social Network?" our system returns "Ben Mezrich", and in this case the inference path should be:

$$\text{The Social Network} \xrightarrow{\text{basedOn}} \text{The Accidental Billionaires} \xrightarrow{\text{author}} \text{Ben Mezrich}$$

In BeamQA [2], we leverage pre-trained language models for their natural language understanding capabilities. A key challenge is that these models are not trained on the property vocabulary of unseen target graphs, which limits their ability to generate valid graph traversal sequences. To overcome this limitation, BeamQA fine-tunes a language model on ground truth (question, path) pairs specifically to generate candidate paths given a question. A scoring algorithm then uses these candidate paths, together with knowledge graph embeddings, to score and rank answers; This ensures that each answer is paired with its corresponding reasoning path.

More recently, we extended this work by integrating large language models (LLMs) such as ChatGPT and LLaMA into an iterative exploration and reasoning framework. Our proposed system, QALLM, alternates between generating the next set of candidate relations, retrieving reachable entities (using KGEs), and then prompting the LLM with the current context to guide its reasoning. By anchoring each reasoning step in the actual graph structure, QALLM leverages the strengths of LLMs, reduces hallucination, and returns a reasoning path alongside every answer.

Post-hoc Explainability. In our previous efforts to make question answering systems provide interpretable reasoning chains, we still rely on components such as knowledge graph embeddings (KGEs), or Graph Neural Networks (GNNs). However, these models function as black boxes and cannot provide interpretations for their link predictions. To address this limitation, we develop post-hoc explainability methods that aim to identify which parts of the input data most influence the model's predictions; typically, the explanation is returned in the form of a subgraph, i.e., a collection of edges or triples ranked by their importance.

We developed XGExplainer [10], an explainability method that enhances the robustness of perturbation-based explainers. Our approach improves the generalizability of existing explainability techniques by decoupling the backbone GNN model from the evaluator that assesses the informativeness of explanatory subgraphs. Experiments demonstrate that XGExplainer consistently improves existing graph explainer performance and outperforms state-of-the-art methods on node and graph classification tasks.

In follow-up work, we address a critical gap in explainability methods for knowledge graph completion by introducing RAWExplainer, a technique that generates interpretable explanations for link predictions made by Graph Neural Networks. When a GNN predicts a missing triple, RAWExplainer identifies a small connected subgraph that explains the prediction by scoring all relevant triples that influenced the model's decision. Unlike existing methods that consider a combinatorial number of possibilities, our approach uses a neural network to parameterize explanation generation. Our approach significantly accelerates computation, provides robustness to distribution shifts through a custom evaluator, and employs constraints to generate connected subgraphs of maximum specified size.

2.3 Impact

My work on interpretable question answering over knowledge graphs has appeared in leading venues such as ICDM 2022, SIGIR 2023, and SDM 2024. This research has produced practical resources through open-source implementations and novel datasets. For example, I supervised Lifan Yu (NYU Shanghai, visiting undergraduate student) in creating CrunchQA [16], a business-domain question answering dataset generated from Crunchbase using template-based question generation. The work has also inspired student research: Ryoji Kubo, whose undergraduate capstone on GNN explainability under my supervision, received the prestigious NYUAD Postgraduate Fellowship to continue collaborating with me on projects such as RAWExplainer; and my PhD student Ola ElKhatib, who is researching LLM-based question answering methods, was awarded the NYUAD-Winton Prize in Science.

3 Heterogeneous Data Integration

3.1 Motivation and Research Challenges

While knowledge graphs provide powerful structured representations of information, real-world entities are inherently dynamic and multimodal. Entities often have visual representations (e.g., satellite images capturing activity in a mall parking lot), evolving or changing properties (e.g., a country's GDP trend or its president), and time series measurements (e.g., daily CO_2 emissions from a power plant). Additionally, such data might originate from diverse sources and formats, including relational data, time series, and multidimensional data. Integrating these modalities enriches entity representations and supports a holistic view for informed decision-making.

To demonstrate the breadth of analytical challenges involving diverse data modalities, my research explores three representative problems: (i) Integrating diverse entity modalities, such as images, (multilingual) text, and graph structure, into unified latent representations to improve knowledge completion, (ii) Adapting question answering frameworks to existing relational databases, and (iii) Investigating how cleaning and imputing temporal data influences downstream analytical tasks.

3.2 Contributions

Multimodal Knowledge Completion. Multimodal graphs (MMGs) are increasingly common in real-world applications. In fact, Wikidata and Wikipedia are inherently multimodal, where entities and articles are associated not only with multilingual text but also with visual content. Multimodal knowledge completion aims to leverage complementary signals from diverse modalities, and the core challenge is to develop techniques that effectively align and integrate these heterogeneous inputs to improve the representation of entities and thereby enhance knowledge completion tasks.

We developed Wiki2Prop [11], a recommender system designed to predict missing relational properties for entities in Wikidata. Wiki2Prop tackles a significant practical challenge in Wikidata knowledge graph completion by assisting editors in efficiently identifying relevant properties among thousands of potential options.

For instance, given an entity such as Marie Curie, missing properties like spouse, award received, or cause of death can often be inferred directly from corresponding textual or visual content. Moreover, relevant information might exist only in some languages or in images. Building on this observation, the system integrates textual and visual embeddings through a neural fusion layer. Compared to existing statistical methods, our approach enhances prediction accuracy and is robust to missing data modalities.

In MLaGA [5], we pursue a different approach by extending LLMs to enhance their ability to reason *with* multimodal graph data. Specifically, MLaGA uses a multimodal encoder that aligns text and image data through graph-aware pre-training. Subsequently, these aligned multimodal features are integrated into the LLM via a lightweight instruction-tuning mechanism. MLaGA significantly improves the effectiveness of LLMs in multimodal graph tasks, such as node classification and link prediction, and demonstrates stronger generalization capabilities across diverse datasets compared to existing approaches.

Question Answering over Tabular Databases. Tabular data is a prevalent and versatile form of structured data, including relational databases, web tables, spreadsheets, and CSV files. Table question-answering systems, such as Text2SQL, commonly convert natural language queries into structured queries for data exploration. However, these solutions typically assume prior knowledge of relevant tables and their schemas, oversimplifying real-world scenarios. In practice, querying large collections of tables lacking a unified or well-documented schema introduces significant challenges involving retrieval and context management.

We address these challenges by adopting a retrieval-augmented generation paradigm that retrieves a subset of relevant information, significantly reducing the context size required for LLM reasoning. Specifically, we introduce DBRAG, an end-to-end table question-answering system with a two-stage retrieval pipeline. First, an encoder-based table index performs a coarse ranking of candidate tables. Second, an LLM refines these top-ranked candidates using a query-driven row context index. To handle potentially large candidate tables during answer generation, DBRAG employs a program-aided chain-of-thought approach integrated with SQL execution tools. This enables the LLM to iteratively execute queries to obtain samples, join tables, and aggregate results, ultimately producing accurate answers from the refined subset of retrieved tables.

Temporal Analytics. Time series data add an important dimension to entities by capturing the dynamic evolution of entity properties through continuous measurements. However, integrating temporal data into knowledge ecosystems presents new infrastructure and data quality challenges. On the infrastructure side, we developed TSM-Bench [9], a comprehensive benchmark to help practitioners select time series database systems for monitoring application requirements. The benchmark features a novel data generation algorithm, a realistic workload, and representative queries inspired by real-world monitoring use cases.

More critically, missing values represent one of the most pervasive data quality issues in temporal data, often appearing as continuous blocks that can severely impact downstream analytics. While time series imputation techniques offer solutions by providing plausible substitutes for missing data, data scientists face critical decisions when selecting appropriate imputation methods, as these choices directly impact the accuracy and reliability of their analyses. To address this challenge, we developed CleanIMP, an evaluation framework that systematically simulates realistic missing-data scenarios, applies a comprehensive suite of imputation algorithms, and benchmarks their impact on downstream applications like forecasting and classification. CleanIMP helps data scientists select imputation strategies that best align with their specific application requirements and dataset characteristics.

3.3 Impact

These research contributions have resulted in publications at top-tier venues, real-world deployments, and new collaborations. Wiki2Prop was published at WWW 2021 and can be activated within Wikidata's interface as an extension that provides editors with automated recommendations to enrich entities by suggesting missing properties. TSM-Bench was published at VLDB 2023; the benchmark requirements were established in consultation with data analysts from the Federal Office for the Environment in Switzerland. We also developed SEER, a web-based companion that demonstrates the benchmark and its data generator [1]. Additionally, DBRAG and MLaGA represent ongoing contributions to database question answering and multimodal reasoning, respectively. This work has fostered multiple productive collaborations on multimodal graph reasoning (Prof. Qiaoyu Tan, NYU Shanghai), time series data quality (Dr. Mourad Khayati, University of Fribourg), and open government data publishing (Dr. Michael Luggen, Swiss Federal Chancellery); with whom we recently developed Cube.Link, an ontology for publishing multidimensional open government data.

4 Future Work

The rapid advancement of artificial intelligence promises transformative opportunities to revolutionize our ability to navigate, interpret, and extract insights from increasingly complex and heterogeneous knowledge ecosystems. Building upon my existing research, I plan to leverage state-of-the-art AI methodologies to develop foundational approaches for structured data completion, exploration, and interpretation. Additionally, I aim to broaden the practical impact of my research through interdisciplinary collaborations and real-world applications. Below, I briefly outline three research directions inspired by my ongoing work.

Scalable Knowledge Graph Models. To address the challenge of incompleteness in very large knowledge graphs, I will investigate transformer-based link prediction models. Inspired by the success of transformer architectures in fields such as natural language processing and computer vision, my objective is to develop foundational models that scale efficiently, and dynamically adapt to evolving knowledge graphs. This research will involve both theoretical advancements and the creation of novel benchmarks. Additionally, I plan to extend these models to incorporate heterogeneous modalities, including textual, visual, and spatio-temporal data. An ongoing interdisciplinary collaboration with Prof. Samer Madanat, integrating graph representations and time series analysis for traffic prediction, exemplifies the practical impact of this work in areas such as traffic control and urban planning.

Conversational AI and Interpretability in Data Science. Generative AI and conversational interfaces have increasingly permeated enterprise environments, impacting not only code development but also analytics and data exploration workflows. Although natural language querying can facilitate user interactions, generative models frequently produce plausible yet incorrect information. My future research will advance beyond simple factual queries toward complex analytical interactions with structured data, including relational databases and knowledge graphs. Specifically, I will investigate methods such as uncertainty quantification, chain-of-thought reasoning, query-driven sampling, and visualization to enhance interpretability and mitigate the risks associated with model hallucinations. Collaborations with industry partners will help comprehend the practical challenges, create realistic benchmarks, and further develop interpretable conversational tools. We have laid out this vision in a collaborative grant proposal with Dr. Stanislav Sobolevsky (Meta), Prof. Muhammad Shafique (NYUAD), and Prof. Siddharth Garg (NYU), submitted to Cisco, titled *"Talking to Networks: A Stochastic Approach to Uncertainty-Aware Spatio-Temporal Knowledge Graph Analytics,"* which is currently under review.

Novel Structured Data Extraction for Societal Impact. Extracting structured data from unstructured textual content offers significant opportunities for building custom knowledge graphs, yet this task remains challenging, even with advanced large language models. My future research will develop targeted relation-extraction pipelines emphasizing specific downstream applications. We will build specialized large language models fine-tuned explicitly for these tasks, combined with human-in-the-loop validation. In an ongoing collaboration with Prof. Etienne Wasmer, we are developing methods to extract mobility trajectories of notable individuals from Wikipedia articles, enabling tasks such as *"Extract all (city, year) pairs from the input article representing where Albert Einstein lived throughout his lifetime."* The datasets generated through this research will enable novel analyses of how influential individuals' movements have historically shaped urban and societal development. Additionally, these datasets will be contributed back to Wikidata for broader community impact.

References

[1] Luca Althaus, Mourad Khayati, Abdelouahab Khelifati, Anton Dignös, Djellel Difallah, and Philippe Cudré-Mauroux. 2024. SEER: An End-to-End Toolkit for Benchmarking Time Series Database Systems in Monitoring Applications. *Proc. VLDB Endow.* 17, 12 (2024), 4361–4364.

[2] Farah Atif, Ola El Khatib, and Djellel Difallah. 2023. BeamQA: Multi-hop knowledge graph question answering with sequence-to-sequence prediction and beam search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 781–790.

[3] Djellel Difallah, Diego Sáez-Trumper, Eriq Augustine, Robert West, and Leila Zia. 2022. Crosslingual Section Title Alignment in Wikipedia. *2022 IEEE International Conference on Big Data (Big Data)* (2022), 5892–5901.

[4] Djellel Difallah. 2025. WikiRAG: Revisiting Wikidata KGC Datasets with Community Updates and Retrieval-Augmented Generation. In *Proceedings of the 31st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '25)*. ACM, Toronto, ON, Canada. (To appear).

[5] Dongzhe Fan, Yi Fang, Jiajin Liu, Djellel Difallah, and Qiaoyu Tan. 2025. MLaGA: Multimodal Large Language and Graph Assistant. *arXiv preprint arXiv:2506.02568* (2025).

[6] Martin Gerlach, Marshall Miller, Rita Ho, Kosta Harlan, and Djellel Difallah. 2021. Multilingual Entity Linking System for Wikipedia with a Machine-in-the-Loop Approach. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*. 3818–3827.

[7] Paul Groth, Elena Simperl, Marieke van Erp, and Denny Vrandecic. 2022. Knowledge Graphs and their Role in the Knowledge Engineering of the 21st Century (Dagstuhl Seminar 22372). *Dagstuhl Reports* 12, 9 (2022), 60–120.

[8] 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. 2022. Knowledge Graphs. *ACM Comput. Surv.* 54, 4 (2022), 71:1–71:37.

[9] Abdelouahab Khelifati, Mourad Khayati, Anton Dignös, Djellel Difallah, and Philippe Cudré-Mauroux. 2023. TSM-Bench: Benchmarking Time Series Database Systems for Monitoring Applications. *Proc. VLDB Endow.* 16 (2023), 3363–3376.

[10] Ryoji Kubo and Djellel Difallah. 2024. XGExplainer: Robust Evaluation-based Explanation for Graph Neural Networks. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 64–72.

[11] Michael Luggen, Julien Audiffren, Djellel Difallah, and Philippe Cudré-Mauroux. 2021. Wiki2Prop: A Multimodal Approach for Predicting Wikidata Properties from Wikipedia. In *WWW*. ACM / IW3C2, 2357–2366.

[12] Nadya Abdel Madjid, Ola El Khatib, Shuang Gao, and Djellel Difallah. 2022. HyperKGQA: Question Answering over Knowledge Graphs using Hyperbolic Representation Learning. *2022 IEEE International Conference on Data Mining (ICDM)* (2022), 309–318.

[13] Natalia Ostapuk, Djellel Difallah, and Philippe Cudré-Mauroux. 2020. SectionLinks: Mapping Orphan Wikidata Entities onto Wikipedia Sections. In *Wikidata@ISWC (CEUR Workshop Proceedings, Vol. 2773)*. CEUR-WS.org.

[14] Natalia Ostapuk, Djellel Difallah, and Philippe Cudré-Mauroux. 2022. ParaGraph: Mapping Wikidata Tail Entities to Wikipedia Paragraphs. In *IEEE Big Data*. IEEE, 6008–6017.

[15] Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. SimKGC: Simple Contrastive Knowledge Graph Completion with Pre-trained Language Models. In *ACL (1)*. Association for Computational Linguistics, 4281–4294.

[16] Lifan Yu, Nadya Abdel Madjid, and Djellel Difallah. 2022. CrunchQA: A Synthetic Dataset for Question Answering over Crunchbase Knowledge Graph. *2022 IEEE International Conference on Big Data (Big Data)* (2022), 4635–4641.