IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5136∼5148
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5136-5148

The first Wikidata-centric agentic artificial intelligence framework for scholarly knowledge

Kemas Rahmat Saleh Wiharja, Ariq Heritsa Maalik, Farrell Habibie Putra Haris, Prati Hutari Gani
School of Computing, Telkom University, Bandung, Indonesia

Article Info

Article history:

Received Jul 14, 2025
Revised Sep 3, 2026
Accepted Sep 19, 2026

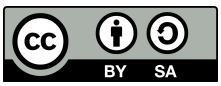
Keywords:

Agentic system
Disambiguation agent
Extraction agent
Knowledge enrichment
Wikidata

ABSTRACT

Large knowledge graphs such as Wikidata contain approximately 1.65 billion structured facts. However, Wikidata still lacks substantial scholarly information, particularly about publications, conferences, and authors. One potential source to close this knowledge gap is scholarly knowledge graphs, such as Google Scholar. This work presents an agentic system capable of extracting academic information, disambiguating authors, validating triples, and inserting verified records into Wikidata. To the best of the knowledge, this is the first Wikidata-centric agentic framework designed to automate the full scholarly knowledge lifecycle—from extraction, disambiguation, and validation to enrichment—using external scholarly data sources. This work evaluates the approach through three tasks: extraction, disambiguation, and Wikidata enrichment. Experiment of this work shows strong results; the workflow successfully extracts relevant metadata, resolves author identities using Google Scholar and ORCID with high confidence levels, and inserts validated data into Wikidata while preventing ambiguous entities from being created.

This is an open access article under the CC BY-SA license.



Corresponding Author:

Kemas Rahmat Saleh Wiharja
School of Computing, Telkom University
Telekomunikasi Street No. 1, Bandung 40287, Indonesia
Email: bagindokemas@telkomuniversity.ac.id

1. INTRODUCTION

Knowledge is something that a person can gain from their own learning experience, or from other learning experience [1]. It could be in the format of information or an understanding of something. In the era of artificial intelligence (AI), knowledge is a critical component that can enable an AI agent to do more than just pattern recognition, and finally can understand, reason, and decide something according to the given context. To store knowledge, one may use a knowledge base (KB), such as Wikidata [2], yet another great ontology (YAGO) [3], never-ending language learner (NELL) [4], Microsoft academic knowledge graph [5], and database systems and logic programming for computer science [6]. In terms of entity and fact volume, Wikidata is considered as one of the largest KB [2], [3]. Until today, Wikidata has more than 100 million entities and around 1.4 billion facts about them [2]. Despite this gigantic size, Wikidata is still far behind Google Scholar (as example) for scholarly/academic data, such as author, co-author, proceeding, and research articles information. Based on the small-scale study for researchers that are affiliated with Telkom University, only 18 researchers can be found in Wikidata page [7], [8], whereas for Google Scholar, +510 researchers found.

Wikidata is an entity-centric knowledge graph: every item—publication, conference, or author—is an entity. Scholarly knowledge graphs such as dblp computer science bibliography (DBLP) and Google

Scholar are publication-centric, organizing everything around the publication [2]. Because Wikidata's focus is structuring machine- and human-editable facts, it holds little scholarly detail. Examining researcher Trong-Minh Hoang [8] as a case study, only two of his six top co-authors were present in Wikidata, as shown in Table 1, illustrating the lack of academic data.

Table 1. The entity coverage of collaborator for Trong - Minh Hoang on Wikidata

Name of co-authors	Wikidata page
Nguyen T. Lang	Not available
Nam-Hoang Nguyen	Not available
Tuan-Anh Pham	https://www.wikidata.org/wiki/Q131009139
Nhu-Ngoc Dao	https://www.wikidata.org/wiki/Q61105681
Nguyen Linh Trung	Not available
Van-Nhan Nguyen	Not available

This gap can be addressed using information extracted from conferences and journals, yielding valid central entities such as authors, paper titles, and affiliations. Extraction approaches have progressed from rule-based and classical machine-learning methods to deep learning (DL), transformer-based models, and more recently, large language model (LLM)-based and agent-based frameworks, of which LLM-driven and knowledge-graph-aware approaches show strong potential for enriching Wikidata. Several approaches enrich Wikidata's knowledge graph. Mihindukulasooriya [9] leveraged DBLP identifiers with the Wikidata reconciliation application programming interface (API) and OpenRefine; Zhang *et al*. [10] integrated 21 million statements via schema alignment and semantic validation (88.46% accuracy, 91.67% F1-score) but depended on external data quality; Rezayi *et al*. [11] proposed enriching knowledge graph embeddings with external text (EDGE), augmenting graph embeddings with text-mined entities at high cost; Peng *et al*. [12] refined taxonomies via LLMs and graph mining (Wikipedia knowledge graph completion (WiKC)); Haussman *et al*. [13] showed domain-specific embeddings (food knowledge graph (FoodKG)) can outperform general models but may not scale; and a recent multi-agent LLM framework [14] targets domain KGs rather than Wikidata directly. These methods advance enrichment but rely on human intervention, variable data quality, or expensive computation.

Adding scholarly data also requires resolving author-name ambiguity first, since author names frequently coincide while titles and affiliations rarely do. Approaches range from rule-based and heuristic methods [10], classical machine learning [11], and graph-based embeddings [12] to bidirectional encoder representations from transformers (BERT)/scientific bidirectional encoder representations from transformers (SciBERT) with graph neural networks (GNNs) [13], and recently, LLMs reasoning over web-scale data—yet most neglect persistent identifiers such as ORCID and sources like Google Scholar; only a few [14] exploit Google Scholar, and none fully exploit ORCID. These advances have not translated into automated Wikidata enrichment: existing efforts remain largely manual, rule-based, or domain-restricted, with no scalable, fully AI-driven framework automating the complete lifecycle—extraction, disambiguation, validation, and insertion—while ensuring only verified triples are added. This research fills that gap with the first Wikidata-centric agentic AI framework for large-scale, automated, ambiguity-aware enrichment.

2. RELATED WORK

This section reviews three areas: information extraction techniques for conferences or journals, author name disambiguation (AND), and Wikidata knowledge graphs as a data source. Goal in this section is to identify limitations that motivate the agentic framework for automated Wikidata enrichment. In each area, this study discusses what the papers do and highlights the weaknesses that the proposed approach can address.

2.1. Information extraction from conferences or journals

Rule- and pattern-based extraction pioneered the field, using manually designed rules, templates, and regular expressions [15], [16]. These approaches achieve high precision in narrow domains but do not scale. Statistical models trained on annotated data [17], [18] generalize better and were widely used from 2005 to 2018, but require heavy feature engineering and are outperformed by DL.

DL, introduced by Hinton *et al*. [19], learns multi-level representations via nonlinear transformations and has advanced speech, vision, and object detection [20]. GNNs model entity relationships and outperform bidirectional long short-term memory (BiLSTM)/convolutional neural network (CNN) for scientific

information extraction [21]. DL removes complex feature engineering and learns sentence-level features automatically via CNNs and recurrent neural network (RNNs) [17], but relies on expensive annotated data, offers limited explainability, and incurs domain-adaptation costs.

Pretraining models—embeddings from language models (ELMo) [22], BERT [23], generative pre-trained transformer (GPT) [24]—substantially improved natural language processing (NLP) via unsupervised pretraining on large corpora. SciBERT [25] addresses scientific text with a domain-specific corpus and vocabulary, improving named entity recognition (NER), relation extraction, and classification even with modest fine-tuning. However, it still relies on task-specific fine-tuning and pipeline architectures, so complexity and error propagation persist. Pre-LLM transformers remain a strong baseline for academic metadata extraction.

The NLP revolution scaled language models from millions [26] to hundreds of billions [27] of parameters on large webtext corpora [28], yielding pretrained LLMs adept at many tasks. With scale came few-shot learning [27], enabling the "pre-train and prompt" paradigm where a well-designed prompt replaces retraining. Prompts may be few-shot or zero-shot [29]; LLMs exhibit strong zero-shot reasoning, understanding complex instructions without intensive domain training, which suits structured extraction from scientific documents. Still, LLMs do not truly learn during inference, demand substantial computation, and the ecosystem is not yet mature enough to replace stable, controlled extraction pipelines.

Knowledge graph is a conceptual model that organizes, integrates, and stores knowledge from various sources so that it can be read and understood by machines [30]. In document representation, traditional methods have limitations, namely the lack of semantic links between words and of word-sequence information. A knowledge graph provides a new perspective and a way of representing knowledge in a document. Documents are represented as entities and their semantic links, not character strings. Wang *et al.* [31] constructs a knowledge graph from the geoscience literature to support scientific reasoning, knowledge exploration, and mineral discovery. Not only that, the knowledge graph can be used to assess whether the outputs of LLMs are more accurate, reliable, and verified [32]. Agent-based planning systems have progressed rapidly [33], driven by LLM advances; however, Dagdelen *et al.* [34] still struggle with complete articles, multimodal data, and human intervention, limiting automation. An automated multi-agent system [35] addresses these shortcomings by extracting structured data from scientific literature.

Overall, information extraction has evolved from rule-based systems and statistical models to DL and LLMs with zero-shot reasoning. Gaps remain in handling multimodal data, managing pipeline error propagation, and ensuring robust semantic validation. The proposed Wikidata-centric agentic framework tries to address these gaps through a task-specific multi-agent architecture that manages the full scholarly knowledge lifecycle.

2.2. Author name disambiguation

Rule-based and heuristic approaches rely on string similarity, initial matching, and affiliation rules to distinguish authors [36]. String similarity handles spelling variants, transliteration, and typos. The initials-matching resolves incomplete first names. Lastly, affiliation rules assume institutional stability over a realistic time window. These approaches are easy to implement and explain, and need no training data, but fail on very common names and degrade on large datasets [36].

Classical machine learning approaches maps publications or publication pairs into feature vectors (co-authorship, venue, and title); it is robust and lightweight but depends on manual features, is sensitive to missing metadata, ignores semantics, and has limited scalability. Graph-based methods model bibliographic data as heterogeneous networks whose nodes represent authors, papers, venues, institutions, and topics, and whose edges capture relations such as writes, co-writes, publishes-in, affiliated-with, and cites [37]. They capture collaboration clusters well but depend on relational-metadata completeness and miss textual semantics. DL integrates structured metadata with pretrained models (BERT/SciBERT) and GNNs [38], but struggles with scarce annotations and imbalance; combining supervised and unsupervised DL is effective while limiting overhead [39]. Recent LLM-based methods reason over web-scale data but mostly neglect ORCID; only a few [40] exploit Google Scholar profiles. Most methods still rely on data completeness and overlook global identifiers, highlighting the need for a more integrated approach to accurate author identification.

2.3. Knowledge graph enrichment

Knowledge graph enrichment (KGE) enhances a knowledge graph by adding new data, increasing completeness, and expanding coverage while maintaining consistency, integrating structured, semi-structured,

and unstructured sources [11]. Embedding-based methods enable semantic integration while probabilistic approaches weigh reliability [41]. Mihindukulasooriya's method [9] enriches Wikidata with DBLP using DBLP IDs, SPARQL protocol and resource description framework query language (SPARQL) metadata extraction, the Wikidata reconciliation API, and OpenRefine, disambiguating via Google Scholar, DOIs, and ORCIDs through a navigable web interface. Few approaches combine NLP-based extraction with human-in-the-loop validation; most are either fully manual or fully automated without quality assurance. The proposed framework integrates AI-based extraction and disambiguation with expert validation before insertion into Wikidata.

2.4. Knowledge graph

A knowledge graph organizes facts into entities, relations, and attributes. Entities can be real-world things or ideas that aren't real. Relations show how these entities are connected, and attributes tell us what they are like [42]. A knowledge graph is formally defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R},$ and $F)$ where $\mathcal{E}$ represents the set of entities, $\mathcal{R}$ denotes the set of relations, and $\mathcal{F}$ signifies the set of factual triples (subject, predicate, and object) [43]. In AI-powered apps, knowledge graphs are often used for semantic search, reasoning, and relational modeling. Constructing a Knowledge graph can be done by hand or with some automation, and it often involves knowledge graph completion (KGC) to fill in the missing links in a triple [42].

2.5. Wikidata

Wikidata is a free, collaborative KB launched by Wikimedia in 2012 to centralize and standardize structured data across all language editions of Wikipedia. Its resource description framework (RDF)-based data model allows multilingual and source-attributed entries represented in triple format. The openness of its editing environment supports real-time updates by a global community [44]. Wikidata is an important part of the Semantic Web because it exports data in RDF, JavaScript object notation (JSON), and other formats under a creative commons zero (CC0) license. Wikidata also lets you use references, qualifiers, and ranks to show conflicting facts. This design encourages a pluralistic way of representing knowledge, which is important for scientific data that comes from many different places. Wikidata is a useful resource for research and data applications [2] because it is well- structured and supports SPARQL.

3. METHOD

The proposed method is shown in Figure 1. The proposed approach is divided into three main stages: the extraction agent, the disambiguation agent, and the insertion mechanism [14], [35]. The first present a formal problem formulation and then detail the design and mathematical foundations of each agent within the pipeline. Operationally, the proposed framework operates as an AI pipeline in which agents sequentially perform i) cross-level information extraction (conference, article, and author), ii) entity disambiguation via external signals, iii) relation extraction between authors and papers, and iv) validation before inserting the knowledge into Wikidata.

3.1. Problem formulation

Let $G_W = (V, E)$ denote a subset of Wikidata knowledge graph [2], [43] for academic domain, where V is the set of entities (e.g., authors, papers, and conferences), whereas the E is the common relationships in academic domain such as authored, published, and written by. Given a list of papers $P = p_1, \ldots, p_n$ that are extracted from a conference that contains the title, the name of authors that authored the paper, the affiliation of each author, and the whole content of the paper, the objective is to automatically insert the novel knowledge item (K_I) from each paper that are not existed in Wikidata to Wikidata. Where K_I is the set of valid novel knowledge item(s) that are extracted from P. To guarantee the level of correctness for the inserted K_I, this study evaluates a random subset of this K_I by using human in the loop mechanism [32]. The resulting knowledge graph is defined in (1).

$$G_{\text{new}} = G_W \cup \left(\sum_{i=1}^{n} K_I \right) \tag{1}$$

3.2. System overview

In this sub section describes in detail the different components of our proposed framework. Figure 1 shows the overall architecture, which comprises three components: two agents operating in series

followed by an enriches. A portable document format (PDF) extractor pulls details from each paper, and a public LLM (OpenAI) structures the results [27], [34]. Each agent is configured with a specialized prompt, hyper-parameters, and domain knowledge. The extraction agent (C) crawls the CEUR Workshop Proceedings at https://ceur-ws.org/, a repository of free open-access computer science proceedings, to locate and retrieve all International Semantic Web Conference (ISWC) papers. The disambiguating agent (D) then resolves author identities against a set of candidates by analysing multiple signals—name, affiliation, co-author, and field consistency. Finally, Wikibase Integrator inserts the new information into Wikidata.

3.3. Extraction agent

This agent extracts three types of information: conference, article, and researcher. For conference information, an OpenAI agent first verifies whether a given uniform resource locator (URL) e.g., https://iswc2025.semanticweb.org/ is a conference website, then passes the verified link to crawl4AI (a dedicated web-crawling framework), which crawls metadata such as committees, accepted papers, and important dates. The output is a JSON object containing the title, location, important dates, topics, URL, year, publisher, organizing committees, and accepted papers (see Figure 2). For article information, a PDF extractor pulls the title, abstract, and author list from each paper as unstructured text, which the OpenAI agent then structures into JSON. For researcher information, no agent is used: this study take each author's name and affiliation and query Google Scholar and ORCID via a Python script—collecting researcher profiles, co-author networks, and publication details from Google Scholar, and the top-5 name-matched researchers from ORCID.

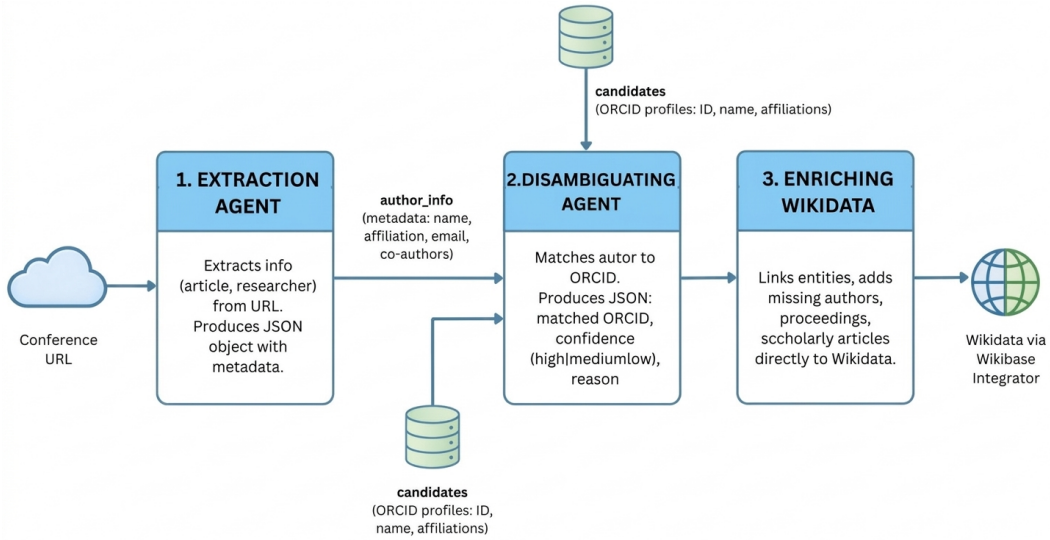


Figure 1. The system architecture of our approach

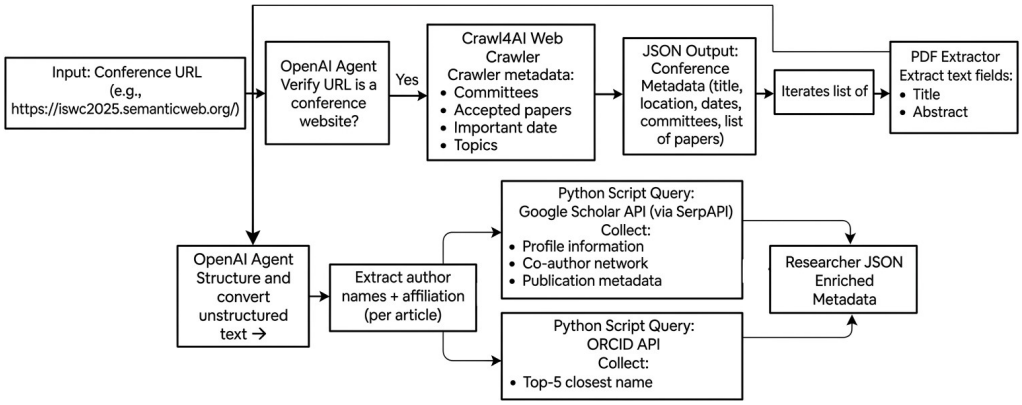


Figure 2. The flowchart of extraction process

3.4. Disambiguating agent

This agent takes two inputs: author_info (name, affiliation, email, and co-authors extracted from a paper) and candidates (a list of ORCID profiles, each with ORCID ID, name, and affiliations). It compares author_info against each candidate using four signals: name match (family and given name, including order and variations); affiliation match (current or past affiliation similar to the paper's); co-author network (overlap between the author's co-authors and those in the candidate's known publications, where available); and field consistency (alignment of affiliations and network with the expected research area, e.g., Semantic Web or knowledge graph). The agent then outputs a JSON record containing the matched ORCID, a confidence level (high, medium, or low), and the reason for the match or non-match.

3.5. Enriching Wikidata

After disambiguation, the system links entities and adds missing authors, proceedings, and scholarly articles to Wikidata via Wikibase Integrator [9], [10]. Each co-author is declared an instance (P31) of "human" (Q5) and populated with available external identifiers: DBLP author ID (P2456), ORCID (P496), open research knowledge graph (ORKG) ID (P10897), Google Scholar ID (P1960), Association for Computing Machinery (ACM) Digital Library author ID (P864), GitHub username (P2037), and X (formerly Twitter) username (P2002). Conference proceedings are asserted as both "conference proceedings" (Q1143604) and "work" (Q3331189), with publisher (P123), language (P407), volume number (P478), DOI (P356), ISBN-13 (P212), and DBLP publication ID (P8978) set accordingly. Scholarly articles are declared instances of "scholarly article" (Q13442814), linked to their venue via "published in" (P1433), and enriched with DBLP ID (P8978), DOI (P356), and page range (P304) where available.

3.6. Explainability and semantic reasoning considerations

Although designed primarily for automated enrichment, the framework inherently supports explainability and semantic reasoning. Each inserted fact is represented explicitly as RDF triples in Wikidata, enabling transparent inspection, provenance tracking, and expert verification, while the disambiguation agent's confidence levels and justification signals (name similarity, affiliation consistency, and co-author networks) make author-resolution decisions interpretable. Ontology alignment is achieved implicitly by grounding all extracted entities and relations in Wikidata's existing data model—predefined classes (e.g., human, scholarly article, and conference proceedings) and properties (e.g., ORCID, Google Scholar ID, and published in)—ensuring interoperability and avoiding schema divergence. The disambiguation-first strategy further enforces semantic consistency before insertion, reducing conflicting or ambiguous entities. Because the agents operate on semantic text representations independent of domain or language, the same workflow extends to multilingual corpora and domain-specific knowledge graphs (e.g., medical or legal knowledge graphs) by substituting only the enrichment module while preserving the extraction and disambiguation modules.

4. RESULTS AND DISCUSSION

In this section presents evaluations for the following task: extraction task which is done by the extraction agent, disambiguation task which is done by our disambiguation agent, and enriching task. For extraction task and disambiguation task, this study compares proposed approaches with respective baselines. As for the enrichment task, this study use ISWC proceeding to enrich Wikidata.

4.1. Extraction task setting

This study use the same dataset as [37]: extracted text from 42 prefaces and websites (22 ISWC, 20 Extended Semantic Web Conference (ESWC)). For each task in Table 2, a domain expert established ground truth by manually extracting facts from the prefaces and official websites, which were then compared against the outputs to compute the precision, recall, and F1-scores. Table 2 also compares proposed method with [37]. The proposed approach achieves perfect extraction for submitted/accepted papers and organizational-role extraction, because ISWC administrators are highly consistent in reporting accepted papers and the committee structure in prefaces is unambiguous. The exceptions are ESWC 2011 and 2019, where the proposed approach hallucinates because these proceedings state accepted papers in an unusual, ambiguous way. For PC-member extraction, the committee structure is inconsistent and has grown increasingly complex since 2019, causing several hallucinations. Important-dates extraction suffers similarly: dates appear as narrative, bullet points, or non-uniform tables (some with many columns), leading to errors such as poster

deadlines being read as paper deadlines. The proposed approach performs better on ESWC than ISWC here, as recent ISWC websites list more important dates. Overall, recurring errors arise mainly when data is presented in inconsistent or highly unstructured formats; future work will add schema-guided extraction templates and prompt refinement to reduce these failures.

Table 2. Evaluation results

Task	Source	Model	ISWC (2002-2023)			ESWC (2004-2023)		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Submitted/accepted papers extraction	Proceedings front matter	GPT-4	0.92	0.87	0.89	0.94	0.95	0.95
		Claude-3	1.00	0.89	0.93	0.90	0.97	0.93
		Proposed approach	1.00	1.00	1.00	0.9	0.9	0.9
Organization role extraction	Proceedings front matter	GPT-4	1.00	1.00	1.00	1.00	1.00	1.00
		Claude-3	1.00	0.96	0.98	1.00	0.92	0.96
		Proposed approach	1.00	1.00	1.00	1.00	1.00	1.00
PC member extraction	Proceedings front matter	GPT-4	1.00	1.00	1.00	1.00	1.00	1.00
		Claude-3	0.98	1.00	0.99	0.99	1.00	0.99
		Proposed approach	0.99	0.98	0.98	0.96	0.97	0.96
Important dates extraction	Website	GPT-4	0.8	0.97	0.88	0.81	0.94	0.87
		Claude-3	0.28	0.92	0.43	0.07	0.81	0.13
		Proposed approach	0.87	0.84	0.86	0.96	0.97	0.97

4.2. Disambiguation task setting

This study uses the AMiner dataset, a large-scale academic knowledge graph maintained by the AMiner team at Tsinghua University and widely used for AND, trend analysis, expert finding, and recommendation. It contains 534,265 publications across 100 blocks, where each block is a common author name shared by multiple distinct researchers; the evaluation set comprises 35,129 publications in 100 blocks. Given publications sharing an author name (e.g., "Wei Zhang"), the task is to cluster those written by the same real-world person, evaluated with pairwise precision, recall, and F1-score against ground-truth labels. This study compares against lazy associative name disambiguation (LAND) [36], a popular disambiguation tool, and Alexandria3k [45], which uses several large language models (LLMs). LAND and proposed approach were evaluated on all 100 blocks, but due to Alexandria3k's heavy resource demands (10 blocks required 7 hours on an A100 40 GB GPU), the comparison with it was limited to the first 10 blocks. Table 3 summarizes the metadata each approach requires.

Table 3. Differences of metadata needed for all three disambiguation approaches

Method	Required metadata	Available in AMiner
LAND [36]	Entity relationships and literals	Knowledge graph triples
Alexandria3k [45]	Titles, venues, and coauthors	Publication metadata
Proposed approach	Candidate profiles and affiliations	Author profile

From Table 4, LAND [36] outperforms proposed approach because AMiner is represented as RDF-like triples that suit an embedding method: its distributed multi-linear gated mechanism textual information (DistMultGateText) model learns the global topology of the co-authorship network, placing entities in a high-dimensional space where "closeness" reflects not just name similarity but latent community structure (shared co-authors, venues, and citation patterns), so it can identify two authors as the same person even when surface metadata is noisy. The proposed LLM-based approach, by contrast, treats disambiguation as semantic matching over publications analyzed in isolation or small batches, lacking this global network view; on AMiner, where many researchers share fields and terminology, semantic analysis alone is insufficient without structural collaboration context. On the 10-block subset, however, the proposed approach outperforms Alexandria3k [45], which disambiguates by pairwise comparison of individual papers, whereas proposed agent aggregates metadata (top co-authors, frequent venues, and affiliations) from all papers into a candidate-profile cluster.

Table 4. The results of AND for the AMiner benchmark

Dataset	LAND [36]			Proposed approach			Alexandria3k [45]		
	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall
A-Miner 100 blocks	0.818	0.708	0.625	0.774	0.453	0.320	NA	NA	NA
A-Miner 10 blocks	0.759	0.724	0.692	0.683	0.320	0.436	0.614	0.297	0.196

4.3. Wikidata enrichment task setting

This study use the Proceeding of ISWC 2024 industry track [9] as the source of documents. For each paper extracts the title, author list, and affiliations; proposed extraction agent identified 34 unique authors, all validated by a domain expert (Table 5), where # denotes the paper number in the ISWC 2024 industry-track proceedings. This study then apply the disambiguation agent to resolve authors with identical or similar names to the same real-world individual (e.g., common-name cases like "John Smith"), using Google Scholar (via SerpAPI) and the ORCID API to identify candidate profiles. Among the 34 authors, only three — Cory Henson, Amit Sheth, and Utkarshani Jaimini — appear in more than one paper, and the agent correctly confirms these as the same individuals. Each author is assigned one status: resolved multiple (candidates found in both sources, best selected by name and affiliation similarity), Google Scholar only (a match only in Google Scholar), not found (no match in either source), or insufficient confidence (candidates found, but none reaching a high enough score). The confidence score (0 – 100) reflects the LLM's assessment based on name similarity, affiliation overlap, and co-author evidence.

Table 5. The result of agent extraction validated by an expert

#	Title	Authors	Expert validation
41	Knowledge representation and engineering for smart diagnosis of cyber physical systems	Ameneh Naghdipour *et al.*	Valid
42	Beyond generic skills: creating-capability centric, company-specific knowledge graphs from job descriptions	Henri Egle Sorotos; Kaan Karakeben; Ahmad Assaf	Valid
43	Causal knowledge graph for scene understanding in autonomous driving	Utkarshani Jaimini; Cory Henson; Amit Sheth	Valid
44	openCypher queries over combined RDF and LPG data in Amazon Neptune	Willem Broekema *et al.*	Valid
45	Causal neuro-symbolic AI for root cause analysis in smart manufacturing	Utkarshani Jaimini; Cory Henson; Amit Sheth; Ramy Harik	Valid
46	Empowering causal machine learning for large-scale manufacturing pipelines with knowledge graphs	Yuxin Zi; Cory Henson; Amit Sheth	Valid
47	A benchmark knowledge graph of driving scenes for knowledge completion tasks	Ruwan Wickramarachchi; Cory Henson; Amit Sheth	Valid
48	Explainability of quality issues in manufacturing: a semantic based approach	Léa Charbonnier; Franco Giustozzi; Julien Saunier; Cecilia Zanni-Merk	Valid
49	Knowledge graph-driven neuro-symbolic system for intelligent document matching	Jans Aasman	Valid
50	Integrating system design information using a self-defining ontology	Stephen Hookway; William Norsworthy Jr.	Valid
51	Assessing the economic impact of the 2024 Baltimore Bridge collapse using RDF/OWL to crosswalk the NAICS, SICS and GICS industry classification systems	Christopher Carrino	Valid

The agent reliably resolves 15 of the 34 authors with confidence $\geq$ 70%; the complete per-author breakdown (disambiguation status, selected source, selected identifier, and confidence score) is provided as Table 6. This confirms that the availability of persistent identifiers and rich profiles (Google Scholar and ORCID) is critical for effective disambiguation, whereas authors with sparse or ambiguous metadata yield insufficient confidence or not found outcomes. The logic is not tied to a specific provider: given comparable metadata (names, affiliations, and co-author networks), the agent extends to services such as DBLP or Semantic Scholar. Finally, this study queried Wikidata via SPARQL to check for existing pages: of the 34 authors, only four already had pages (Stefan Schlobach, Utkarshani Jaimini, Amit Sheth, and Ramy Harik). Using Wikibase Integrator, this study created 30 new author pages, adding each author's name, affiliation, and authored paper(s) as the notable-work attribute. A domain expert then validated each page—checking identifier values, adding valid references where proposed agent's values were incorrect, and verifying other

attributes for authors lacking ORCID or Google Scholar accounts; the full expert-validation record (Wikidata QID and ORCID validity status per author) is provided as Table 7. Of the 30 created authors, only one (Q137375959) was removed by a Wikidata administrator under the notability policy; all others were successfully created, with linked papers inserted from Google Scholar. As a controlled proof-of-concept limited to expert-validated samples, large-scale coverage, throughput, comparative benchmarking, inter-annotator agreement, and editor-facing usability were not assessed and are left to future work.

Table 6. Disambiguation results

Name	Disambiguation status	Selected source	Selected ID	Confidence (%)
Ameneh Naghdipour	resolved_multiple	google_scholar	pUM6OHcAAAAJ	90
Benno Kruit	resolved_multiple	google_scholar	iKI-htQAAAAJ	90
Jieying Chen	insufficient_confidence			
Peter Kruizinga	not_found			
Godfried Webers	not_found			
Stefan Schlobach	resolved_multiple	orcid	0000-0002-3282-1597	90
Henri Egle Sorotos	not_found			
Kaan Karakeben	not_found			
Ahmad Assaf	resolved_multiple	google_scholar	gJ-kXmkAAAAJ	90
Utkarshani Jaimini	resolved_multiple	orcid	0000-0002-1168-0684	90
Cory Henson	resolved_multiple	orcid	0000-0003-3875-3705	70
Amit Sheth	resolved_multiple	orcid	0000-0002-0021-5293	90
Willem Broekema	not_found			
Mohamed Elzarei	insufficient_confidence			
Ora Lassila	resolved_multiple	google_scholar	udmYLRAAAAAJ	90
Carlos Manuel Lopez Enriquez	not_found			
Marcin Neyman	not_found			
Florian Schmedding	not_found			
Michael Schmidt	insufficient_confidence			
Andreas Steigmiller	not_found			
Geo Varkey	not_found			
Gregory Todd Williams	google_scholar_only	google_scholar	OQhcHmoAAAAJ	90
Amanda Xiang	not_found			
Ramy Harik	resolved_multiple	google_scholar	RlD4AOcAAAAJ	90
Yuxin Zi	resolved_multiple	google_scholar	ajOv3NcAAAAJ	90
Ruwan Wickramarachchi	resolved_multiple	google_scholar	p-IrBkAAAAJ	90
Léa Charbonnier	resolved_multiple	google_scholar	-k4zju8AAAAJ	90
Franco Giustozzi	resolved_multiple	orcid	0000-0002-2709-2625	90
Julien Saunier	resolved_multiple	orcid	0000-0002-7385-4395	70
Cecilia Zanni-Merk	resolved_multiple	google_scholar	flwP3NIAAAAJ	90
Jans Aasman	insufficient_confidence			
Stephen Hookway	not_found			
William Norsworthy Jr.	not_found			
Christopher Carrino	not_found			

4.4. Adaptability to other knowledge bases

Although proposed implementation targets Wikidata, the proposed agent-based framework is not limited to a single KB. The proposed approach consists of three modules as follow: knowledge/information extraction, disambiguation, validation, and enrichment/insertion. This study implements the three modules as modular components. As result, this allows the enrichment module to be adapted to other KBs such as DBpedia or OpenCyc. In such cases, proposed insertion module can be replaced with KB-specific APIs or SPARQL endpoints, while the extraction and disambiguation agents remain unchanged. Schema alignment can be achieved by mapping extracted entities and relations to the target ontology, for example aligning scholarly articles and authors to DBpedia classes or OpenCyc concepts. Furthermore, the disambiguation-first strategy and confidence-aware validation are KB-agnostic and can support consistency enforcement across different KBs. Demonstrating downstream societal applications that consume the enriched knowledge (e.g., semantic search, question answering, or educational tools) is beyond the scope of this proof-of-concept and is therefore not addressed in this paper. This design enables the framework to generalize beyond Wikidata while preserving semantic integrity.

Table 7. Expert validation of agent-suggested ORCIDs

Name	Wikidata QID	ORCID suggested by agent	Expert validation
Ameneh Naghdipour	Q137375907	0000-0001-9961-4819	ORCID is not valid
Benno Kruit	Q137375908	0000-0002-0228-4823	valid
Jieying Chen	Q86946859	0000-0001-5847-3853	ORCID is not valid
Peter Kruizinga	Q137375910		valid
Godfried Webers Q137375911		valid	
Henri Egle Sorotos Q137375912		valid	
Kaan Karakeben Q137375913		valid	
Ahmad Assaf	Q137375914	0000-0003-0415-8705	ORCID is not valid
Cory Henson	Q137375948	0000-0003-3875-3705	valid
Willem Broekema	Q137375950		valid
Mohamed Elzarei	Q137375951	0000-0002-2198-004X	ORCID is not valid
Ora Lassila	Q137375952	0000-0002-0001-9137	valid
Carlos Manuel Lopez Enriquez	Q137375953		valid
Marcin Neyman	Q137375954		valid
Florian Schmedding	Q137375955		valid
Michael Schmidt	Q137375956	0000-0002-4492-9730	valid
Andreas Steigmiller	Q137375958		valid
Geo Varkey	Q137375959		remove by the Wikidata administrator
Gregory Todd Williams	Q137375961		valid
Amanda Xiang	Q137375962		valid
Yuxin Zi	Q137375963	0009-0006-9694-1359	valid
Ruwan Wickramarachchi	Q137375965	0000-0001-7004-8702	ORCID is not valid
Léa Charbonnier	Q137375966	0009-0002-0902-9822	valid
Franco Giustozzi	Q137375967	0000-0002-2709-2625	valid
Julien Saunier	Q137375968	0000-0002-7385-4395	valid
Cecilia Zanni-Merk	Q137375969	0000-0002-5189-9154	valid
Jans Aasman	Q137375970	0000-0001-6977-6710	valid
Stephen Hookway	Q137375972		valid
William Norsworthy Jr.	Q137375973		valid
Christopher Carrino	Q137375974		valid

5. CONCLUSION

This paper presented the design, development, and evaluation of an agentic framework for enriching Wikidata in the domain of conference and publications. The proposed agentic framework consists of extraction agent, disambiguation agent, and Wikidata enrichment module that implements using modular design principles therefore, this design allows our approach to be reused for enriching other types of KBs. This study performed a complete evaluation, ranging from extraction task, disambiguation task, and Wikidata enrichment task. The experimental results demonstrate robust performance across the scholarly knowledge lifecycle: the extraction agent achieves consistently high precision and recall across heterogeneous conference sources, while the disambiguation-first strategy effectively prevents ambiguous or conflicting author identities from being introduced into Wikidata. The proposed analysis further indicates that the effectiveness of LLM-based disambiguation is strongly influenced by the availability of persistent external identifiers, such as ORCID and Google Scholar profiles, underscoring the importance of integrating authoritative external knowledge into agentic AI systems. Moreover, the modular architecture allows the framework to be adapted to alternative KBs, such as DBpedia and OpenCyc, by modifying only the enrichment module while preserving the core extraction and disambiguation agents. This study also acknowledge ethical considerations in scalable knowledge-graph enrichment, including potential bias caused by uneven scholarly representation across sources, the need for source reliability when using external metadata, and transparency in validation; ambiguous cases therefore require human review in workflow of this study, and future iterations will explore explicit bias-mitigation strategies and improved auditability.

6. FUTURE WORK

As future work, this study plans to extend proposed framework to accommodate other additional scholarly data sources (e.g., DBLP, Semantic Scholar, and Microsoft Academic Knowledge Graph). To mitigate the extraction-related hallucinations that arise when conference metadata is presented in inconsistent or

unstructured formats, this study will incorporate schema-guided extraction templates, and prompt refinement. This study also intends to incorporate the embedding approach and transformer-based models to our agentic framework to increase the performance and robustness of our disambiguation agent. In addition, this study plan to extend the extraction pipeline with multilingual support so the framework can process non-English scholarly content. For the enrichment module, this study plan to attach more references to the attributes created for each author page, thereby increasing verifiability. This study will also involve a larger number of domain experts to validate the extended framework. Finally, this study will explore confusion-matrix-style plots, real-time enrichment and visualization of structural impact—such as before-and-after changes in Wikidata coverage and graph density—once large-scale deployment is conducted.

FUNDING INFORMATION

This work was supported by the internal grant of Telkom University, grant number 166/LIT06/PPM-LIT/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Kemas Rahmat Saleh Wiharja	✓	✓		✓	✓	✓			✓	✓	✓	✓	✓	
Ariq Heritsa Maalik		✓	✓			✓		✓	✓	✓	✓	✓		
Farrell Habibie Putra Haris	✓		✓	✓		✓			✓		✓		✓	
Prati Hutari Gani					✓		✓	✓		✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The full dataset that support the findings of this study will be released in GitHub at https://github.com/Farrellhrs/WikiEnrich following a 6-month embargo from the date of publication to allow for the commercialization of research findings.

REFERENCES

[1] T. Sapkota, "Teachers', principals' and university-based teacher educators' multiple conceptions of teachers' professional development," *Frontiers in Education*, vol. 10, Sep. 2025, doi: 10.3389/feduc.2025.1554740.

[2] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014, doi: 10.1145/2629489.

[3] F. M. Suchanek, M. Alam, T. Bonald, L. Chen, P.-H. Paris, and J. Soria, "YAGO 4.5: a large and clean knowledge base with a rich taxonomy," in *47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2024, pp. 131–140, doi: 10.1145/3626772.3657876.

[4] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell, "Toward an architecture for never-ending language learning," *24th AAAI Conference on Artificial Intelligence*, vol. 24, no. 1, pp. 1306–1313, Jul. 2010, doi: 10.1609/aaai.v24i1.7519.

[5] J. Priem, H. Piwowar, and R. Orr, "OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts," in *26th International Conference on Science, Technology and Innovation Indicators (STI 2022)*, 2022, pp. 1–8, doi: 10.5281/zenodo.6936227.

[6] A. Rosenfeld, "Is DBLP a good computer science journals database?," *Computer*, vol. 56, no. 3, pp. 101–108, Mar. 2023, doi: 10.1109/MC.2022.3181977.

[7] K. R. S. Wiharja, "SPARQL query for Telkom University researchers in Wikidata," drive.google.com. Accessed: Jun. 28, 2026. [Online]. Available: https://surl.li/gbbbuf

[8] T.-M. Hoang, "Trong-Minh Hoang," Google Scholar. Accessed: Jun. 28, 2026. [Online]. Available: https://scholar.google.com/citations?user=j_I0HGMAAAAJ&hl=en

[9] N. Mihindukulasooriya, "DBLP to Wikidata: populating scholarly articles in Wikidata," in *ISWC 2024 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice*, 2024.

[10] B. Zhang, F. Ilievski, and P. Szekely, "Enriching Wikidata with linked open data," in *Wikidata'22: Wikidata workshop at 21st International Semantic Web Conference (ISWC 2022)*, 2022, pp. 1–16.

[11] S. Rezayi, H. Zhao, S. Kim, R. Rossi, N. Lipka, and S. Li, "Edge: enriching knowledge graph embeddings with external text," in *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2767–2776, doi: 10.18653/v1/2021.naacl-main.221.

[12] Y. Peng, T. Bonald, and M. Alam, "Refining Wikidata taxonomy using large language models," in *33rd ACM International Conference on Information and Knowledge Management*, Oct. 2024, pp. 5395–5399, doi: 10.1145/3627673.3679156.

[13] S. Haussmann *et al*., "FoodKG: a semantics-driven knowledge graph for food recommendation," in *The Semantic Web – ISWC 2019*, Cham: Springer, 2019, pp. 146–162, doi: 10.1007/978-3-030-30796-7_10.

[14] Y. Lu, Y. Lu, W. Wu, X. Zhao, R. Peng, and J. Wang, "KARMA: leveraging multi-agent LLMs for automated knowledge graph enrichment," in *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[15] L.-T. Wu, J.-R. Lin, S. Leng, J.-L. Li, and Z.-Z. Hu, "Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web," *Automation in Construction*, vol. 135, Mar. 2022, doi: 10.1016/j.autcon.2021.104108.

[16] S. Zhou, B. Yu, A. Sun, C. Long, J. Li, and J. Sun, "A survey on neural open information extraction: current status and future directions," in *Thirty-First International Joint Conference on Artificial Intelligence-Survey Track*, 2022.

[17] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, and Z. Wang, "A survey of information extraction based on deep learning," *Applied Sciences*, vol. 12, no. 19, Sep. 2022, doi: 10.3390/app12199691.

[18] L. Pai *et al*., "A survey on open information extraction from rule-based model to large language model," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 9586–9608, doi: 10.18653/v1/2024.findings-emnlp.560.

[19] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[21] N. Wang, T. Chen, C. Ren, and H. Wang, "Document-level relation extraction with multi-layer heterogeneous graph attention network," *Engineering Applications of Artificial Intelligence*, vol. 123, Aug. 2023, doi: 10.1016/j.engappai.2023.106212.

[22] M. Peters *et al*., "Deep contextualized word representations," in *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[24] G. Yenduri *et al*., "GPT (generative pre-trained transformer)—a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, vol. 12, pp. 54608–54649, 2024, doi: 10.1109/ACCESS.2024.3389497.

[25] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: a pretrained language model for scientific text," in *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3613–3618, doi: 10.18653/v1/D19-1371.

[26] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," in *5th International Conference on Learning Representations (ICLR 2017)*, 2017.

[27] T. Brown *et al*., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

[28] J. Dodge *et al*., "Documenting large webtext corpora: a case study on the colossal clean crawled corpus," in *2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 1286–1305, doi: 10.18653/v1/2021.emnlp-main.98.

[29] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *36th International Conference on Neural Information Processing Systems*, 2022, pp. 22199–22213.

[30] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge graphs: opportunities and challenges," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13071–13102, Nov. 2023, doi: 10.1007/s10462-023-10465-9.

[31] C. Wang, X. Ma, J. Chen, and J. Chen, "Information extraction and knowledge graph construction from geoscience literature," *Computers & Geosciences*, vol. 112, pp. 112–120, Mar. 2018, doi: 10.1016/j.cageo.2017.12.007.

[32] S. Tsaneva, D. Dessì, F. Osborne, and M. Sabou, "Knowledge graph validation by integrating LLMs and human-in-the-loop," *Information Processing & Management*, vol. 62, no. 5, Sep. 2025, doi: 10.1016/j.ipm.2025.104145.

[33] M. M. Eissa, "Protection techniques with renewable resources and smart grids—A survey," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 1645–1667, Dec. 2015, doi: 10.1016/j.rser.2015.08.031.

[34] J. Dagdelen *et al*., "Structured information extraction from scientific text with large language models," *Nature Communications*, vol. 15, no. 1, Feb. 2024, doi: 10.1038/s41467-024-45563-x.

[35] R. Odobesku *et al*., "Agent-based multimodal information extraction for nanomaterials," *npj Computational Materials*, vol. 11, no. 1, Jun. 2025, doi: 10.1038/s41524-025-01674-7.

[36] C. A. D'Angelo, C. Giuffrida, and G. Abramo, "A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 2, pp. 257–269, Feb. 2011, doi: 10.1002/asi.21460.

[37] C. Rastogi, P. Agarwal, and S. Singh, "Exploring graph based approaches for author name disambiguation," Dec. 2023, *arXiv:2312.08388*.

[38] Z. Boukhers and N. B. Asundi, "Whois? deep author name disambiguation using bibliographic data," in *Linking Theory and Practice of Digital Libraries*, Cham, Switzerland: Springer Nature, 2022, pp. 201–215, doi: 10.1007/978-3-031-16802-4_16.
[39] J. Gong *et al.*, "MORE: toward improving author name disambiguation in academic knowledge graphs," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 1, pp. 37–50, Jan. 2024, doi: 10.1007/s13042-022-01686-5.
[40] P. Sancheti, K. Karlapalem, and K. Vemuri, "LLM driven web profile extraction for identical names," in *Companion Proceedings of the ACM Web Conference 2024*, May 2024, pp. 1616–1625, doi: 10.1145/3589335.3651946.
[41] Z. Huang, H. Zhang, C. Hao, H. Yang, and H. Wu, "A cross-domain transfer learning model for author name disambiguation on heterogeneous graph with pretrained language model," *Knowledge-Based Systems*, vol. 305, Dec. 2024, doi: 10.1016/j.knosys.2024.112624.
[42] M. Habiburahman, K. Wiharja, and M. Fikriansyah, "Beyond benchmarks: assessing knowledge graph completion methods on non-benchmark employee data," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, Jul. 2024, pp. 28–33, doi: 10.1109/ICoDSA62899.2024.10652136.
[43] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, Feb. 2022, doi: 10.1109/TNNLS.2021.3070843.
[44] K. Guo, D. Diefenbach, A. Gourru, and C. Gravier, "Wikidata as a seed for web extraction," in *Proceedings of the ACM Web Conference 2023*, Apr. 2023, pp. 2402–2411, doi: 10.1145/3543507.3583236.
[45] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, "A brief survey of automatic methods for author name disambiguation," *ACM SIGMOD Record*, vol. 41, no. 2, pp. 15–26, Aug. 2012, doi: 10.1145/2350036.2350040.

BIOGRAPHIES OF AUTHORS



Kemas Rahmat Saleh Wiharja holds a Ph.D. in Computing Science, from University of Aberdeen, United Kingdom. He also receive his B.Sc. and M.Sc. from Telkom University, and Universitas Gadjah Mada, respectively. He is currently an associate professor of Symbolic Artificial Intelligence in Telkom University, Indonesia. He is an active researcher in the domain of knowledge graph, and semantic web. Until 2026, he has published 68 papers with 134 citations. He can be contacted at email: bagindokemas@telkomuniversity.ac.id.



Ariq Heritsa Maalik is a generative AI engineer at Amdocs Limited, where he builds LLM-powered agentic systems for the telecommunications industry. He received his S.Si.D. degree in Data Science, School of Computing, Telkom University, Bandung, Indonesia, graduating cum laude in 2025. His research interests include knowledge graph enrichment, linked data, agentic systems, and large language models, with a broader focus on big data analytics and artificial intelligence. He can be contacted at email: ariqhm@gmail.com.



Farrell Habibie Putra Haris is an AI/ML engineer for iOS App Development. He has contributed to industry projects through internships at PT. Infomedia Nusantara, Transtrack, and the Apple Developer Academy, Tangerang. He received his S.Si.D. degree in Data Science, School of Computing, Telkom University, Bandung, Indonesia, graduating with a GPA of 3.82 in 2026. His research interests include large language models, knowledge graphs, computer vision, time-series forecasting, information retrieval, and artificial intelligence. He can be contacted at email: farrellhrs@gmail.com.



Prati Hutari Gani is a lecturer at Telkom University. She received her master's degree in Informatics Engineering from the Institut Teknologi Bandung. She is currently pursuing a Ph.D. in Informatics at Telkom University, where her research focuses on knowledge graphs, particularly their application in intelligent information systems and artificial intelligence. Her areas of expertise include database systems, data warehousing, artificial intelligence, knowledge graphs, data visualization, and social media analytics. She is actively involved in teaching, research, and community service, with numerous scientific publications and competitive research grants. She can be contacted at email: pratihutarigani@telkomuniversity.ac.id.