# Docs2KG: A Human-LLM Collaborative Approach to Unified Knowledge Graph Construction from Heterogeneous Documents

Qiang Sun
pascal.sun@research.uwa.edu.au
The University of Western Australia
Perth, WA, Australia

Yuanyi Luo
luoyy@stu.hit.edu.cn
Harbin Institute of Technology
Harbin, China

Wenxiao Zhang
wenxiao.zhang@research.uwa.edu.au
The University of Western Australia
Perth, WA, Australia

Sirui Li
sirui.li@uwa.edu.au
The University of Western Australia
Perth, WA, Australia

Jichunyang Li
jichunyang.li@uwa.edu.au
The University of Western Australia
Perth, WA, Australia

Kai Niu
kai.niu@research.uwa.edu.au
The University of Western Australia
Perth, WA, Australia

Xiangrui Kong
xiangrui.kong@research.uwa.edu.au
The University of Western Australia
Perth, WA, Australia

Wei Liu
wei.liu@uwa.edu.au
The University of Western Australia
Perth, WA, Australia

## ABSTRACT

Even for a conservative estimate, over 80% of enterprise data resides in unstructured documents spanning diverse formats and modalities, posing significant challenges for knowledge extraction, association and representation. Although large language models (LLMs) have shown promising capabilities in text processing, their limitations in maintaining factual accuracy and document provenance necessitate complementary approaches. Knowledge graphs offer a structured framework for grounding and verifying information [6], yet existing methods struggle to construct high-quality KGs from heterogeneous data sources. To address this issue, we present **Docs2KG**, a modular framework to build high-quality knowledge graphs from diverse unstructured documents. Docs2KG first employs state-of-the-art document processing techniques to extract textual content, tabular data, and figures. The extracted information is then unified into a multifaceted knowledge graph with three aspects: (1) a <u>Layout KG</u> capturing document structural hierarchies, (2) a <u>Metadata KG</u> preserving document properties, and (3) a <u>Semantic KG</u> representing domain-specific entities and relationships. To ensure flexibility and extensibility, **Docs2KG** supports multiple construction paradigms for Semantic KG: ontology-based approaches, hybrid NLP pipelines with LLM verification, and LLM-guided ontology generation. The framework also allows seamless integration of specialized models for named entity recognition, event extraction, and causal relationship identification to enhance semantic coverage and accuracy. A key feature of **Docs2KG** is its human-in-the-loop verification interface, enabling iterative quality assessment and refinement of the resulting knowledge graphs. **Docs2KG** is openly available at https://docs2kg.ai4wa.com, with the aim of advancing knowledge graph construction research and accelerating enterprise applications through high-quality knowledge graph construction.

## KEYWORDS

Unstructured Data, Heterogeneous Data, Knowledge Graph


*Under Review, , Perth, WA*

ACM ISBN xxx-xxx-xxx
https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Document-centric knowledge management faces significant challenges as unstructured documents proliferate across enterprises in various formats (e.g., words, web pages, PDFs) and modalities (e.g., text, tables, images), with these heterogeneous sources accounting for over 80% of corporate data lakes [7]. The absence of standardized structure in these documents, coupled with the diverse formats and implicit semantic relationships among modalities, makes it particularly challenging to extract, integrate, and utilize the valuable knowledge embedded within them for downstream applications.

While Large Language Models (LLMs) demonstrate remarkable capabilities in natural language understanding and generation, they face critical challenges in enterprise applications due to hallucination and the inability to effectively ground responses in source documents. Knowledge graphs address these limitations by providing a structured representation that explicitly captures semantic relationships and maintains document provenance, enabling reliable fact verification and context-aware reasoning through Retrieval Augmented Generation (RAG) [4]. This necessitates the development of a robust documents-to-knowledge-graph pipeline that can effectively process heterogeneous documents and construct comprehensive knowledge representations.

The aimed pipeline typically comprises two critical stages: document digitization and knowledge graph construction. Although document digitization—particularly for scanned PDFs—has historically been challenging, requiring sophisticated layout analysis and OCR techniques, recent advancements in this area have significantly improved extraction accuracy for both text and rich elements like tables and figures. However, the knowledge graph construction stage remains a significant bottleneck. Traditional approaches require extensive manual annotation (bottom-up) or subject domain expertises (SMEs) for ontology construction (top-down), making full automation impractical. Although recent attempts leverage LLMs for automated knowledge graph construction, they face limitations in output quality and domain generalizability, particularly in specialized fields. The integration of these two stages into a robust, end-to-end pipeline thus presents unique challenges, primarily stemming from the knowledge graph construction phase rather than document digitization at the current stage.

In this work, we present a modularized pipeline for knowledge graph construction from unstructured documents. The pipeline first utilizes existing document digitization technologies (e.g., Docling [7], MinerU [8]) to extract text, tables, and figures. These extracted elements are then processed through our knowledge graph construction framework, which builds knowledge graphs comprising three aspects: Layout KG, Metadata KG, and Semantic KG. While Layout and Metadata KGs follow well-defined construction rules, the Semantic KG construction adapts to different scenarios: (1) for domains with established ontologies, we employ ontology-based construction with LLM prompting, (2) for scenarios with predefined entity or relation lists, we implement traditional NLP-based extraction with LLM verification, and (3) for cases without prior knowledge structures, we use LLM to generate an initial domain ontology based on domain descriptions. For domain-specific applications, ontologies and entity lists can be bootstrapped from public annotation datasets. The framework also supports the integration of specialized models for named entity recognition, event extraction, and causal relation extraction. Finally, the constructed knowledge graphs undergo human verification through an annotation interface, enabling quality assurance and model improvement.

## 2 RELATED WORK

***Document Digitization*** Previous work in document digitization can be broadly categorized into two streams based on input formats. The first stream addresses native digital documents (e.g., web pages, office documents, emails, generated PDFs) that are inherently machine-readable and can be processed using conventional parsing techniques [7]. The second stream addresses scanned PDF documents, which present unique challenges due to their image-based nature, necessitating sophisticated pipelines for document understanding that incorporate layout analysis, optical character recognition (OCR), table recognition, etc.

Retrieving private knowledge from unstructured documents to augment LLMs has emerged as a critical research direction due to its potential impact. This task is currently bottlenecked by challenges in scanned PDF digitization, which has prompted significant advancements in the past six months. State-of-the-art systems such as IBM's Docling [7] and Shanghai AI Lab's MinerU [8] have pioneered dual-path architectures that apply lightweight parsing to machine-readable documents while processing scanned documents through advanced deep learning pipelines incorporating OCR, layout detection, and table extraction. The widespread adoption of these systems is evidenced by their substantial GitHub popularity, with Docling and MinerU garnering 14.1k and 21.2k stars respectively. Recent advances in specialized models have further enhanced the capabilities of these systems, with Da et al. achieving 96.2 mAP @ IOU for layout detection and Wei et al. attaining a 0.972 F1 score for OCR. These advances enable reliable conversion of documents into semi-structured machine-readable formats (e.g., JSON, Markdown), facilitating downstream applications such as knowledge graph construction.

***Knowledge Graph Construction*** has traditionally followed two approaches: **top-down**, where domain experts first develop a comprehensive ontology to guide the construction process, and **bottom-up**, where the ontology emerges from manual entity and relation annotations. Both approaches heavily depend on human expertise, requiring deep domain knowledge for ontology design and substantial manual effort for annotation.

Prior to the advent of Large Language Models (LLMs), knowledge graph construction typically prioritized human-driven ontology development, supplemented by specialized deep learning based models (e.g., domain-specific Named Entity Recognition) to partially automate the annotation process. However, this approach faced significant scalability challenges, particularly in ontology development, which required extensive cross-domain expert communication to achieve high-quality knowledge representation.

Recent approaches have explored using LLMs as automated agents to replace human involvement in both annotation and ontology development processes. While frameworks like Langchain [1] offer automated entity and relation extraction, these bottom-up approaches often yield knowledge graphs of insufficient quality without subsequent ontological refinement. Recent research has increasingly recognized the critical role of ontologies in improving knowledge graph construction. For example, SPIRES [1] achieves enhanced performance by strategically incorporating predefined ontologies into prompts to guide LLM-based extraction. Similarly, Text2KGBench [5] proposes a comprehensive framework that combines ontology-based prompt generation, LLM-driven knowledge extraction, and post-processing steps for entity/relation refinement. LLMs4OL [3] focuses on using LLMs to generate ontologies as a preliminary step in the construction process. These works demonstrate that while LLMs offer promising capabilities for automated knowledge graph construction, their effectiveness is substantially improved when guided by well-defined ontological frameworks.

To achieve automatic domain-agnostic knowledge graph construction, several fundamental challenges persist. Large Language Models (LLMs) often lack the deep domain-specific understanding necessary for accurate knowledge representation. Additionally, prompt-based extraction methods introduce non-deterministic behavior, as the quality of extracted knowledge varies significantly based on prompt design. Furthermore, evaluating the quality of automatically constructed knowledge graphs presents its own challenges, as standard metrics are limited and often require validation through the performance of downstream tasks [10].

## 3 SYSTEM DESIGN

The **Docs2KG** framework transforms documents into high quality knowledge graphs through a multi-stage pipeline (Figure 1). The system performs 1 metadata extraction and MetadataKG construction, followed by *Dual-Path Document Digitization* using **Docling** or **MinerU** tools to generate standardized outputs (Markdown for texts, JSON for tables, and image files for figures). The digitized

[1] https://python.langchain.com/v0.1/docs/use_cases/graph/constructing/

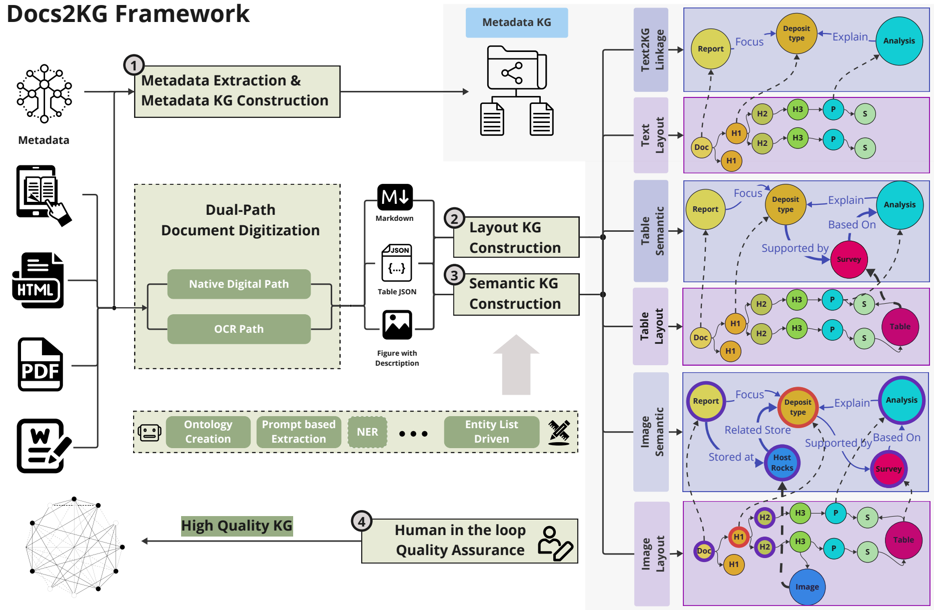


**Figure 1: Docs2KG Framework Design: Multifaceted Knowledge Graph (MetadataKG, LayoutKG, SemanticKG) Construction followed by Human-in-the-loop Quality Assurance. SemanticKG adopts an extensible, modular pipeline design.**

content undergoes (2) Layout and (3) Semantic KG construction, with SemanticKG featuring a modular pipeline architecture. (4) A human-in-the-loop quality assurance process refines the resulting multifaceted multimodal knowledge graph.

***Multifaceted KG Construction:*** **Metadata KG** Our MetadataKG schema formalizes document metadata, which inherently exists in tabular formats across document management systems or within documents themselves, as a directed property graph $G_{metadata} = (V, E, \Phi_v, \Phi_e)$, where vertices $V$ represent document and enumerated metadata entities, and edges $E$ denote their relationships. Document entities ($v_d \in V_d$) incorporate standard properties such as filenames, alongside other properties like temporal $\phi_t \in \Phi_v$ (e.g., creation date) or spatial $\phi_s \in \Phi_v$ (e.g., polygons) properties where applicable. Enumerated metadata fields including document types and authorship information are represented as distinct entity types ($V_{type}, V_{author} \subset V$) and linked to documents via typed edges ($e_t, e_a \in E$), facilitating efficient metadata-driven retrieval and reasoning.

**Layout KG** Our LayoutKG schema captures document structural hierarchies, which mimic human visual information processing patterns, as a directed property graph $G_{layout} = (V, E, \Phi_v, \Phi_e)$, where vertices $V$ represent textual elements of different granularities (e.g., chapters, sections, paragraphs). These vertices are connected through edges $e \in E$ that encode structural relationships ('has-child', 'before', 'after'), enabling hierarchy-aware document traversal and retrieval.

**Semantic KG** Our SemanticKG schema formalizes domain knowledge and cross-modal relationships as a directed property graph $G_{semantic} = (V, E, \Phi_v, \Phi_e)$, where vertices $V$ represent domain concepts (e.g., geological formations, tectonic events) and multimodal content (e.g., tables, figures, and their textual descriptions). These vertices are connected through edges $e \in E$ that encode semantic relationships ('explains', 'coexists', 'causes'), enabling both knowledge grounding against established geological concepts and hypothesis investigation through novel relationship discovery.

The implementation complexity varies across our three *faceted* knowledge graphs. MetadataKG and LayoutKG utilize straightforward rule-based mapping: metadata fields become graph properties, while document elements (sections, paragraphs, tables, figures) are linked through hierarchical and sequential relationships. SemanticKG construction, however, adapts to resource availability through three main pathways: (1) ontology-driven extraction using domain-specific patterns and LLM prompting when ontologies exist [1], (2) entity-list-driven extraction, where traditional NLP methods extract entities from texts based on predefined entity lists, with LLM verification ensuring domain-appropriate extractions, and (3) LLM-assisted dynamic ontology generation from document content and domain context when no prior ontologies exist. This flexible framework enables future integration of specialized Named Entity Recognition (NER), Event, or Causal Event extractors.

***Human Verification:*** The evaluation of KG construction quality faces two key challenges: (1) no standardized metrics exist for evaluating KG quality constructed from given text, and (2) no established thresholds define sufficient KG quality for practical use. While downstream tasks could serve as evaluation methods, this approach is both time-consuming and difficult to scale, potentially hindering the development of more efficient KG construction methods. Instead of tackling these challenges, we propose a pragmatic human-in-the-loop approach enabling high quality KG construction across domains, an example is shown in Figure 2 and 3. Automatically constructed KGs will be presented through an annotation interface where domain experts can modify both entities and relations (instances and types).

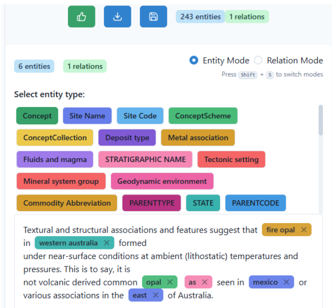


Figure 2: KG edit Interface

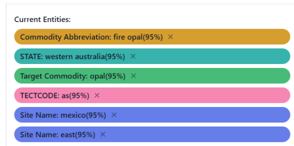


Figure 3: extracted entities

To evaluate KG quality, we propose two simple metric: **Human-LLM Opinion Distance** $D = \alpha \frac{|E_h \triangle E_a|}{|E_h \cup E_a|} + \beta \frac{|R_h \triangle R_a|}{|R_h \cup R_a|}$ between original and expert-edited KGs, where $E_h$, $E_a$ represent entity sets, $R_h$, $R_a$ represent relation sets, $\triangle$ denotes symmetric difference, and weights $\alpha + \beta = 1$. To quantify each method's contribution, we define **Contribution Factor** $C_i = \frac{D_{without\_i} - D_{combined}}{D_{without\_i} + \epsilon}$, where $D_{without\_i}$ is the score without method $i$ and $D_{combined}$ is the score with all methods. A lower $D$ indicates better KG quality as it shows fewer differences from expert edits, while a higher $C$ indicates greater contribution as it reflects larger quality degradation when the method is removed. $\epsilon$ is a small positive constant added to prevent division by zero. The annotation interface is free accessible via https://docs2kg.kaiaperth.com/, where you can import, save, edit, and export the unified KG and automatically generate the evaluation metrics.

## 4 CASE STUDY

The Western Australian Mineral WAMEX (WAMEX) database from Geological Survey of Western Australia (GSWA)[2] contains over 100,000 geological reports spanning the past century, primarily in PDF format (both scanned and digital). Similar to most enterprise systems, WAMEX maintains well-structured tabular metadata for these reports, including creation dates and geospatial information[3]. We also extract 215,147 *point of interest* entities cover 67 entities types from its transactional databases, particularly from GSWA MINEDex[4]. Additionally, unlike most enterprises, GSWA realized the value of ontologies early on and has a valuable domain ontology under active development[5].

We first establish MetadataKG and LayoutKG through rule-based approaches. For SemanticKG construction, we employ a two-stage process: (1) Entity list-driven extraction followed by Phi3.5[6] as LLM verification agent, and (2) Ontology-based extraction [1] using Phi3.5 as KG construction agent. We also explored automatic ontology creation using Phi3.5, followed by the approach in [1]. Evaluation metrics are shown in Table 1.

[2]https://wamex.dmp.wa.gov.au/Wamex
[3]https://dasc.dmirs.wa.gov.au/home?productAlias=MinExpRepWAMEX
[4]https://minedex.dmirs.wa.gov.au/
[5]https://vocabulary.gswa.kurrawong.ai/
[6]https://huggingface.co/microsoft/Phi-3.5-mini-instruct

**Table 1: KG construction evaluation**

| Method | $D$ | $C$ |
|---|---|---|
| Combined | 0.25 | - |
| Entity list | - | 0.29 |
| Ontology | - | 0.23 |
| Auto-ontology | 0.45 | - |

## 5 CONCLUSION

We present **Docs2KG**, a human-LLM collaborative framework for constructing high-quality unified knowledge graphs from heterogeneous enterprise documents. Our approach combines human expertise with LLM-based automation to enhance KG generation while reducing manual effort. The unified multifaceted knowledge graph includes MetadataKG, LayoutKG, and SemanticKG. We propose evaluation metrics to measure the gap between human and automatic pipelines, enabling quick bottleneck identification and targeted improvements. Our WAMEX case study demonstrates the effectiveness of this collaboration, achieving high quality (D=0.25). The high contribution score of the entity list-driven approach suggests that enterprises can achieve decent quality KGs by extracting *point of interest entities* from their existing transactional databases, which aligns with intuition as most valuable business domain knowledge is already modeled within these databases. This unified framework provides enterprises a practical solution to transform heterogeneous document repositories into high-quality, structured knowledge graphs for various downstream applications.

## REFERENCES

[1] J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, Peter N Robinson, and Christopher J Mungall. 2024. Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. Bioinformatics 40, 3 (Feb. 2024). https://doi.org/10.1093/bioinformatics/btae104
[2] Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. 2023. Vision Grid Transformer for Document Layout Analysis. arXiv:2308.14978 [cs.CV] https://arxiv.org/abs/2308.14978
[3] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2023. LLMs4OL: Large Language Models for Ontology Learning. arXiv:2307.16648 [cs.AI] https://arxiv.org/abs/2307.16648
[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33 (2020), 9459–9474.
[5] Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F. Enguix, and Kusum Lata. 2023. Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text. arXiv:2308.02357 [cs.CL] https://arxiv.org/abs/2308.02357
[6] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. arXiv:2307.07697 [cs.CL] https://arxiv.org/abs/2307.07697
[7] Deep Search Team. 2024. Docling Technical Report. Technical Report. https://doi.org/10.48550/arXiv.2408.09869 arXiv:2408.09869
[8] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. MinerU: An Open-Source Solution for Precise Document Content Extraction. arXiv:2409.18839 [cs.CV] https://arxiv.org/abs/2409.18839
[9] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model. arXiv:2409.01704 [cs.CV] https://arxiv.org/abs/2409.01704
[10] Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. arXiv:2305.13168 [cs.CL] https://arxiv.org/abs/2305.13168