Docs2KG: Unified Knowledge Graph Construction from Heterogeneous Documents Assisted by Large Language Models

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, 80% of enterprise data reside in unstructured files, stored in data lakes that accommodate heterogeneous formats. Classical search engines can no longer meet information seeking needs, especially when the task is to browse and explore for insight formulation. In other words, there are no obvious search keywords to use. Knowledge graphs, due to their natural visual appeals that reduce the human cognitive load, become the winning candidate for heterogeneous data integration and knowledge representation. In this paper, we introduce Docs2KG, a novel framework designed to extract multimodal information from diverse and heterogeneous unstructured documents, including emails, web pages, PDF files, and Excel files. Dynamically generates a unified knowledge graph that represents the extracted key information, Docs2KG enables efficient querying and exploration of document data lakes. Unlike existing approaches that focus on domain-specific data sources or pre-designed schemas, Docs2KG offers a flexible and extensible solution that can adapt to various document structures and content types. The proposed framework unifies data processing supporting a multitude of downstream tasks with improved domain interpretability. Docs2KG is publicly accessible at https://docs2kg.ai4wa.com, and a demonstration video is available at https://docs2kg.ai4wa.com/Video.

KEYWORDS

Unstructured Data, Heterogeneous Data, Knowledge Graph

1 INTRODUCTION

The most valuable enterprise knowledge reside in unstructured documents of heterogeneous formats, taking up at least 80% of the corporate data lakes. It is crucial to extract meaningful information [7] by integrating these data, while maintaining references to the origin for Retrieval Augmented Generation (RAG) [5] to reduce hallucination. Taking the healthcare industry as an example, patient records often exist in various formats such as handwritten clinical notes, discharge letters, email communication between clinicians, and medical images. Without data integration, it is impossible to provide a consolidated assessment. Many existing works [6, 7] are designed to target a single data source, such as scanned documents or PDF files. However, in real-world applications, particularly within domain-specific knowledge areas, data are heterogeneous, unstructured, and diverse [8]. To perform document-wide semantic parsing and layout analysis from heterogeneous unstructured documents, we face three key challenges:

- The extraction of multimodal data (incl. tables, texts, images, and figures) from a diverse range of formats.
- Integrating modality-specific information extraction models into one unified framework.
- Meaningful representation of data semantic with references to the source.

In this research, we propose using Knowledge Graphs as a unified representation to allow dynamic integration of entities extracted from each modality, including layout entities to maintain references to the source. The end goal of knowledge graph construction is faciliated through our proposed **Docs2KG** system to address the above challenges.The data formats that Docs2KG can handle include emails, web pages, PDF files, and Excel files. The extracted multimodal information, merged as a unified KG, allows for dynamic and automatic update based on document structure and content, which can be modified and extended to allow human-in-the-loop. It enables researchers and domain experts to pose structural and semantic queries such as *“Show me all documents and their components related to events that occurred in the years 2011 and 2021.“*. This

Under Review, , Perth, WA

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

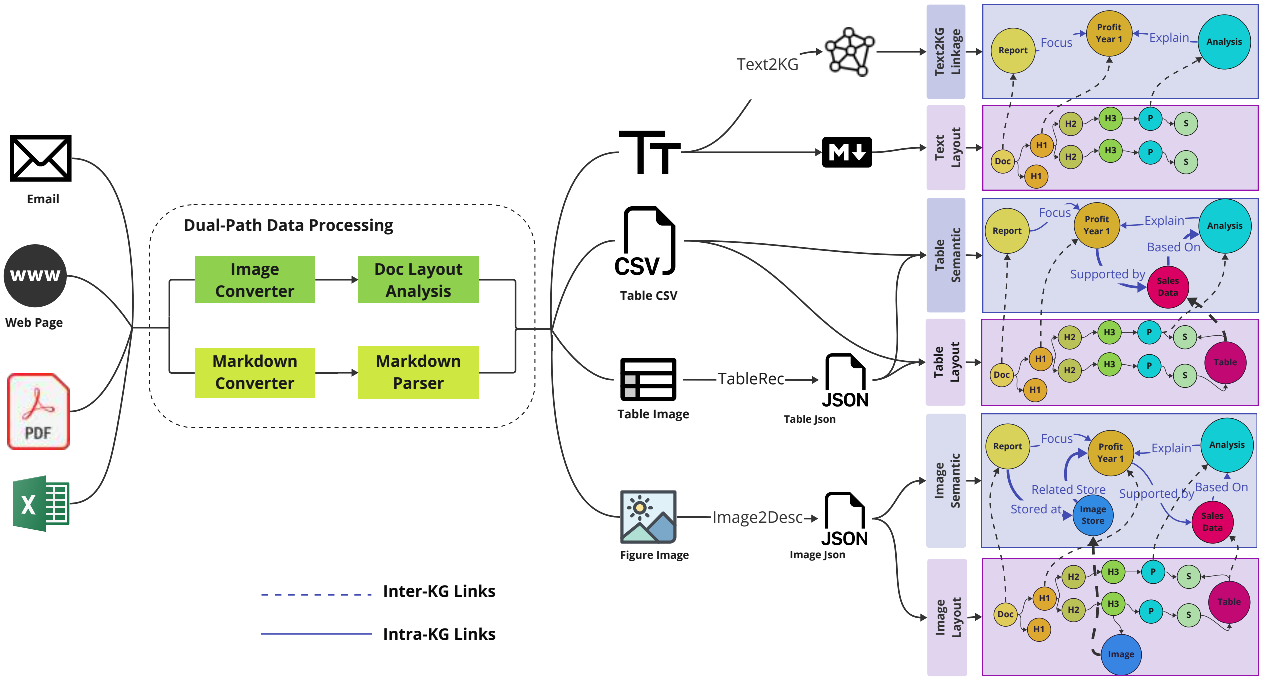


Figure 1: Architecture Design for Docs2KG

capability can dramatically reduce the time, effort, and resources required navigating through large collections of unstructured documents. Moreover, Docs2KG unified document processing through *a dual path strategy* which effectively combined deep learning computer vision based document layout analysis with mark-down structured document parsing to maximise its document type coverage. The KG generated by Docs2KG can be used to facilitate many real-world applications, such as reducing the risk of outdated knowledge and hallucination of language language models to achieve knowledge-grounded retrieval augmented generation.

2 RELATED WORK

There have been several efforts to construct KGs to facilitate the discovery of relevant information within specific fields. Most of these efforts [1, 4, 9] have focused on extracting information from text. For example, Connected Papers [1] is a tool designed to help researchers and academics to find and explore relevant academic papers. It creates a citation network of papers for a given search paper, allowing users to see connections and discover influential works in their field. This visualisation aids in the literature search in a broader context assisting in finding seminal works and new directions worth investigation. Another example is the work by Kannan et al. [3], who built a multimodal KG that extracts text, diagrams, and source code from scientific literature in the field of Deep Learning.

Our framework, Docs2KG, differs from these approaches by specifically targeting at heterogeneous unstructured documents rather than just scientific publications. While their schema is predesigned for specific domains, such as deep learning architectures, ours is dynamic and automatically generated based on the document structure. Additionally, Docs2KG can be modified and extended as needed, making it more adaptable to various types of unstructured data.

3 DOCS2KG FRAMEWORK

The architecture of Docs2KG is shown in Figure 1, which is designed to take asinput a set of heterogeneous and unstructured documents, including emails, web pages, PDF files and Excel files. Docs2KG involves two main stages: dual-path data processing and multimodal unified KG construction. The *dual-path data processing* stage segments the input documents into textual content, images, and tables. The *multimodal unified KG construction stage* integrates the processed information with structural and semantic relationships.

After alignment, the resulting multimodal KG is stored in a Neo4j[1] graph database, allowing storage of the extracted information a triple store for efficient querying and intuitive visualisation. All code and documentation are available online[2]. The code is designed to be modualised, other graph databases can be used to replace Neo4j for graph data storage and retrieval. The following sections detail the two key stages of Docs2KG.

3.1 Dual-Path Data Processing

In Figure 1, we categorise the input documents into two types based on the easy of extracting their layout information. For example, web pages (HTML) are organised using a tree structure, enabling straightforward conversion to Markdown or JSON. In contrast,

[1] https://neo4j.com/
[2] https://docs2kg.ai4wa.com/

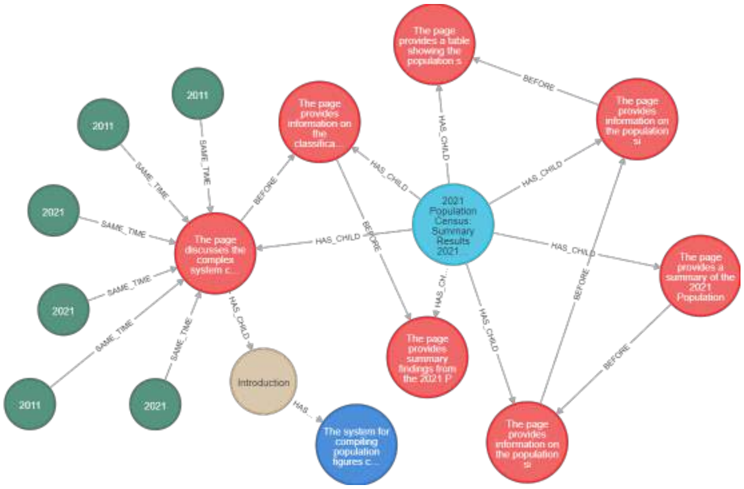

The page provides a table showing the population s
BEFORE
2011
2011
SAME_TIME
SAME_TIME
The page provides information on the classifica...
HAS_CHILD
The page provides information on the population si
HAS_CHILD
HAS_CHILD
BEFORE
2021
SAME_TIME
The page discusses the complex system c...
HAS_CHILD
2021 Population Census: Summary Results 2021
BEFORE
HAS_CHILD
The page provides a summary of the 2021 Population
2021
SAME_TIME
SAME_TIME
SAME_TIME
HAS_CHILD
HAS_CH...
HAS_CHILD
BEFORE
BEFORE
2011
2021
Introduction
The page provides summary findings from the 2021 P
HAS...
The system for compiling population figures c...
The page provides information on the population si

Figure 2: A demo graph of query "Show me all documents and their components related to events that occurred in the years 2011 and 2021." by combining a PDF file and an Excel file. The PDF file contains information about the population size and structure of Hong Kong from 2011 to 2021. The Excel file contain records of the population census from 2021 to 2023. (Cyan indicates the PDF document; Green is for Excel file; Red for PDF page; Khaki for header; ocean blue for paragraph)

PDF files and Excel files with extensive descriptive information pose significant challenges for layout detection and transformation into semi-structured format. To address the above challenges, we propose a dual path document processing strategy. The *Image Converter* path is a generic approach that uses deep learning models trained for document layout analysis; the *Markdown Converter* path is to convert documents to markdown format and use an XML/HTML query language such as XPath. All four types of documents can be converted into images and take advantage of the document layout analysis to segment into texts, images, and tables with bounding boxes. We will not provide details on how these are achieved; please refer to our publications on PDF form data analysis [11]. For markdown document parsing, we have developed four independent parsers to process different document types:

- **PDF parsing:** Based on the meta information provided by the PDF file, we can determine whether to feed it to the **Markdown Converter** or **Image Converter**. For scanned PDF files, the only path is through trained document layout analysis models, while generated PDF files can be parsed or segmented to extract images, tables, texts with bounding box information.
- **Web page parsing:** We use a popular Python library, `BeautifulSoup` [2], for efficient HTML parsing. Texts are extracted using `markdownify` [10]. Images are identified via the `<img>` tag, tables via the `<table>` tag. The original document tree structure of the HTML page is retained as alayout knowledge graph.
- **Excel parsing:** Using the Python library `pandas`, Excel files are loaded and data are extracted from each worksheet. The extracted data is then converted into images via *imgkit*, and then go through the **Image Convertor** path. For complex structured Excel worksheets, they can converted to PDF files first, to follow the PDF processing pipeline.
- **Email parsing:** We assume emails are in .eml format. The Python library `email` is then used to segment messages into plain text, HTML, and attachments. Text and HTML sections of the emails can then be processed similarly to web pages, while attachments are handled by appropriate tools based on their formats, such as PDF or Excel parsers.

By combining parsers and document segmentation models, Doc2KG can parse different heterogeneous and unstructured documents for subsequent integration into a unified KG. The modualised approach we are taking allow for flexible configuration and combination of the processing modules to optimise computation resource usage.

3.2 Multimodal Unified Knowledge Graph Construction

After the first stage, our proposed Docs2KG unifies the parsed information into a multimodal KG containing structural (hierarchical and spatial) and semantic information.

We categorise relationships of our multimodal KG into two primary types: intra-modal relationship and inter-modal relationship.

Intra-modal relationships construction: Intra-modal relationships include structural relationships at the title level and paragraph level, and semantic relationships at the sentence level. The intra-modal relationships can be expressed as:

$$G^{(\alpha,\beta)} = (h_\alpha, r, t_\beta), \alpha \neq \beta \in \{T, P, S\} \quad (1)$$

where the G represents a smallest unit sub-graph in our multimodal KG. α and β represent different modalities from text source, containing text (T), paragraph (P), and sentence (S). The notation (h_α, r, t_β) denotes the construction method between two nodes, where h_α (the head entity) points towards t_β (the tail entity). r denotes the relationship, expressed with structural or semantic information:

- **Structural relationships:** 'has-child', 'before' and 'after'.

- **Semantic relationships:** 'same time', 'focus', 'supported by', 'explain'.

Inter-modal relationships construction: We use semantic relationships to express the relationships between different modalities. It is because the intra-modal hierarchical and spacial relationships already provide a clear relationship direction. The inter-modal relationships can be expressed as:

$$G^{(S,M)} = (h_S, r, t_M), M \in \{Table, Figure\} \qquad (2)$$

where G represents a smallest unit sub-graph. S denotes sentences, such as table captions. M denotes tables and figures. r is the semantic relationship between them: 'explain' and 'same-time'.

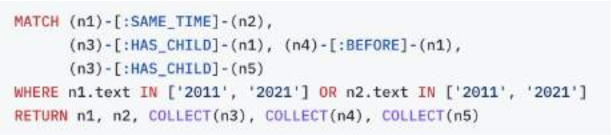
4 DEMONSTRATION

In our demonstration, we first focus on how our multimodal KG can be utilised to perform data-driven analysis through a graph querying demo. Subsequently, we demonstrate how the KG can support one of the most important applications of large language models, RAG. In our RAG demo, nodes and relationships are embedded and subjected to a similarity search to identify anchor nodes. These nodes are then expanded via multi-hop queries to retrieve relevant information, thereby augmenting the prompt to respond to the query.

4.1 Knowledge Graph Query

We selected one PDF file and one Excel file for the demo. The PDF file contains information about the population size and structure of Hong Kong from 2011 to 2021. The Excel file contain records of the population census from 2021 to 2023, including mid-year population data categorised by age group and sex.

Meaningful insights cannot be derived from either the Excel file or the PDF file alone. We parsed and integrated the PDF file and the Excel file through Docs2KG. The data were extracted into figures, tables, and text, and merged into a single KG. To extract relevant information, we used the query shown in Figure 3. The returned graph is in Figure 2 where green bubbles and red bubbles represent the information extracted from Excel and PDF files, respectively. Based on the visualisation, we can observe that the introduction section (the Khaki coloured node) of the PDF document references several events occurring in both 2011 and 2021. For more information about this demo, please refer to our demo video https://docs2kg.ai4wa.com/Video/.

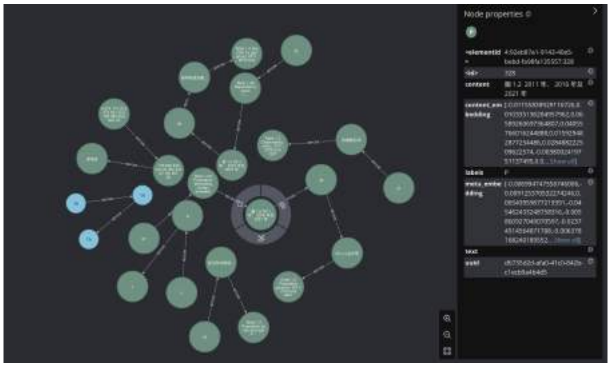


```
MATCH (n1)-[:SAME_TIME]-(n2),
      (n3)-[:HAS_CHILD]-(n1), (n4)-[:BEFORE]-(n1),
      (n3)-[:HAS_CHILD]-(n5)
WHERE n1.text IN ['2011', '2021'] OR n2.text IN ['2011', '2021']
RETURN n1, n2, COLLECT(n3), COLLECT(n4), COLLECT(n5)
```

Figure 3: The Cypher Query to answer "Show me all documents and their components related to events that occurred in the years 2011 and 2021."

4.2 Semantic and Structural Proximity-Based Information Retrieval

To enhance the performance of large language models, the RAG approach suggests integrating more relevant information directly into the prompt. In the context of our multimodal knowledge graph, 'relevance' refers to the proximity of nodes, which can be either semantic or structural. Specifically, relevant nodes are those that can be reached within a limited number of hops in the knowledge graph.



Node properties

<elementId>	[illegible]
<id>	[illegible]
content	[illegible]
content_embedding	[illegible]
labels	P
meta_embedding	[illegible]
text	
uuid	[illegible]

Figure 4: Retrieved relevant semantic and structural nodes for query "I want to know all the population information from 2011 to 2021" by combining the same files referenced in Section 4.1. (Green indicates <p> tag; Blue for <tr> tag.)

Based on this, consider the same query in above demonstration: "I want to know all the population information from 2011 to 2021". Initially, all nodes within the knowledge graph are embedded using an embedding model. The same model is used to embed the query. The query embedding is then utilized to retrieve relevant text chunks, figures, and tables through semantic similarity search. The top-k semantically relevant nodes will be selected as anchor nodes to retrieve the n-hop semantic and structural relevant nodes, there by augmenting the prompts as shown in Figure 3. We can see the tables regarding the population information from 2011 to 2021 are retrieved. For additional details regarding this demonstration, please refer to our demo video at https://docs2kg.ai4wa.com/Video/ or our codes.

5 CONCLUSION

In this paper, we have addressed the limitations of existing multimodal KG construction methods by proposing an open-source framework, Docs2KG. Unlike previous approaches that either focus solely on images or rely on an existing KG to link images, our framework considers more realistic scenarios across all domains. Docs2KG effectively handles the diversity and heterogeneity of raw data in various unstructured formats, such as web pages, emails, PDF files, and Excel files. By integrating these diverse data sources into a unified KG and incorporating both semantic and structural information, Docs2KG enables a more comprehensive and accurate representation of knowledge. This facilitates a wide range of real-world applications, improving the utility and robustness of KGs in diverse domains.

REFERENCES

[1] Alex Tarnavsky Eitan, Eddie Smolyansky, Itay Knaan Harpaz, and S Perets. 2021. Connected papers. S. l (2021).

[2] Gábor László Hajba. 2018. Using Beautiful Soup. Apress, Berkeley, CA, 41–96. https://doi.org/10.1007/978-1-4842-3925-4_3

[3] Amar Viswanathan Kannan, Dmitriy Fradkin, Ioannis Akrotirianakis, Tugba Kulahcioglu, Arquimedes Canedo, Aditi Roy, Shih-Yuan Yu, Malawade Arnav, and Mohammad Abdullah Al Faruque. 2020. Multimodal knowledge graph for deep learning papers and code. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 3417–3420.

[4] Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. In The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA (AAAI Technical Report, Vol. WS-17). AAAI Press. http://aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15129

[5] 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.

[6] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. DocBank: A Benchmark Dataset for Document Layout Analysis. In Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 949–960. https://doi.org/10.18653/V1/2020.COLING-MAIN.82

[7] Siwen Luo, Yihao Ding, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. Doc-GCN: Heterogeneous Graph Convolutional Networks for Document Layout Analysis. In Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 2906–2916. https://aclanthology.org/2022.coling-1.256

[8] Mohammed Maree and Mohammed Belkhatir. 2015. Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies. Knowledge-Based Systems 73 (2015), 199–211.

[9] Anderson Rossanez and Júlio Cesar dos Reis. 2019. Generating Knowledge Graphs from Scientific Literature of Degenerative Diseases. In Proceedings of the 4th International Workshop on Semantics-Powered Data Mining and Analytics co-located with the 18th International Semantic Web Conference (ISWC 2019), Aukland, New Zealand, October 27, 2019 (CEUR Workshop Proceedings, Vol. 2427), Zhe He, Jiang Bian, Cui Tao, and Rui Zhang (Eds.). CEUR-WS.org, 12–23. https://ceur-ws.org/Vol-2427/SEPDA_2019_paper_8.pdf

[10] Matthew Tretter. 2024. Markdownify: A library for converting HTML to Markdown. https://github.com/matthewwithanm/python-markdownify Accessed: 2024-05-18.

[11] Haolin Wu, Tim French, Wei Liu, and Melinda Hodkiewicz. 2022. Automatic semantic knowledge extraction from electronic forms. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 0, 0 (2022), 1748006X221098272. https://doi.org/10.1177/1748006X221098272 arXiv:https://doi.org/10.1177/1748006X221098272