Developing Relation Types of Cryptocurrency Anti-Money Laundering Knowledge Graph

Min-Yuh Day *
Graduate Institute of Information Management
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

Pei-Tz Chiu
Graduate Institute of Information Management
National Taipei University
New Taipei City, Taiwan
s711036103@gm.ntpu.edu.tw

Yung-Wei Teng
Graduate Institute of Information Management
National Taipei University
New Taipei City, Taiwan
s711036115@gm.ntpu.edu.tw

Chao-Lin Liu
Department of Computer Science
National Chengchi University
Taipei City, Taiwan
chaolin@g.nccu.edu.tw

***Abstract*—Anti-money laundering involving cryptocurrencies has become a popular research topic in recent years. Moreover, constructing a knowledge graph of cryptocurrency anti-money laundering in a small sample of judgments to prevent cryptocurrency money laundering has become an essential issue for an improved understanding of the relationship between crime patterns and emerging financial technologies. The research method of this study is that we conducted a named entity recognition task and identified the key relation types to construct a cryptocurrency anti-money laundering knowledge graph (KG). Accordingly, we developed the "Judicial7," a key relation type for cryptocurrency anti-money laundering KG. The contribution of this study is that the proposed "Judical7" relation types of cryptocurrency anti-money laundering KG can be applied to construct a legal knowledge graph.**

***Keywords*—knowledge graph, named entity recognition, relation type, cryptocurrency anti-money laundering, artificial intelligence.**

I. INTRODUCTION

The term of knowledge graph (KG) was first introduced by Google in 2012. Google Search engine utilizes the Google KG as knowledge base to provide user information. To date, KG has become a prevailing form for representing knowledge. For downstream applications, knowledge-aware applications incorporate recommendation systems, question answering systems, and understand natural language, among others [3].

Surveys, such as that conducted by Teichmann and Falker [5], have shown that how cryptocurrencies are used to launder money. Money launderers continue to abuse cryptocurrencies, such as Bitcoin, as a means for committing financial crimes. For these reasons, we realize that understanding the connection between crime patterns and financial technologies is critical.

Day [7] indicated that the construction of cryptocurrency anti-money laundering KG can explain the connection between crime patterns and financial technology [7]. To understand the contribution of KG in the cryptocurrency domain, the current study focuses on identifying the relation types of cryptocurrency-related crimes.

The remainder of this paper is structured as follows. Section 2 presents a literature review on KG overview and domain-specific KG. Section 3 establishes the research methodology and design. Section 4 shows the experimental results and discussion. Lastly, Section 5 presents the conclusions of this research.

II. RELATED RESEARCH

A. Overview of KG

KG, as a new form of knowledge representation, is represented by the resource description framework (RDF), the triples like "head entity-relationship-tail entity," which also denotes <*h*, *r*, *t*> [10].

One of the major tasks for constructing KG is knowledge extraction. Abu-Salih [11] explained that knowledge extraction includes entity and relation-levels. The former incorporates named entity recognition (NER), named entity disambiguation (NED) and named entity linking (NEL). For the latter, relation-level knowledge extraction is divided into global-based and local-based relation extraction (RE).

The second stage of KG construction is knowledge reasoning. To date, knowledge reasoning via KG has become a popular research topic because it can obtain new knowledge from existing data [12]. As Chen, et al. [12] indicated that knowledge reasoning task can be completed based on logic rules, distributed representation, and neural network. Furthermore, KG completion (KGC) is a task that cannot be omitted in knowledge reasoning for knowledge graph construction and related applications because it aims to complete the KG structure by predicting the missing entities or relationships in KG and mining unknown facts [10].

Common knowledge graphs in the field include WordNet, FreeBase, and YAGO. Statistics of general KG data sets when originally released are shown in Table I [3, 7, 13].

TABLE I. STATISTICS OF GENERAL KNOWLEDGE GRAPH DATASETS

Dataset	# Entities	# Relations
WN11	38,696	11
WN18	40,943	18
WN18RR	40,943	11
FB13	75,043	13
FB15k-237	14,541	237
FB40K	39,528	1,336
YAGO3-10	123,182	37

* Corresponding author.

978-1-6654-6603-5/22/$31.00 ©2022 IEEE
DOI 10.1109/IRI54793.2022.00031

B. Domain-specific KG

In recent years, there has been an increase in the literature on domain-specific KG. We reviewed some studies in various fields.

In the medical domain, Li et al. [1] presented a DeepKG, which is a deep learning-based functionality, to obtain significant knowledge from the coronavirus disease 2019 (COVID-19) full-text literature. The DeepKG adopts the Auto TransX, which is a novel automated machine learning (AutoML)-based method for embedding knowledge and training a model to extract triplets data, including subject, predicate, and object (SPO), to complete KG. The model generates high-quality KG on biological data with 7980 entities and 43760 triplets knowledge. The KG and extensive experiments constructed by Li et al. [1] showed evidence of the effectiveness of downstream tasks, such as question answering systems and generating candidate drug list predicted by Auto TransX.

In the education field, Li, et al. [2] introduced a multi-modal educational KG (MEduKG). First, Li et al. [2] first fine-tuned a bidirectional encoder representation from transformers (BERT) model to extract valuable information in the education domain. Second, the combination of the bidirectional long short-term memory and a conditional random field (BiLSTM-CRF) model was adopted to recognize educational entities. Third, information of entities was incorporated into BERT to predict the educational relationship between entities. Lastly, a speech-fusion method was adopted to complete MEduKG.

In the biomedical field, Xu et al. [4] introduced the PubMed KG (PKG). PubMed is an important textual knowledge resource in the medical field. However, extracting useful concepts is relatively ambiguous and difficult. To solve this problem, Xu et al. [4] established the PKG by extracting biological entities from 29 million PubMed abstracts, which utilizes the high-performance deep learning method called BERT for biomedical text mining (BioBERT), integrates research funding data through the National Institutes of Health (NIH) ExPORTER, and eliminates the author's name to construct the PKG by collecting the author's affiliation history and educational background from ORCID. In addition, fine-grained affiliation data are identified from MapAffil. By integrating these reliable multi-source data, relationships among biological named entities, authors, articles, affiliations, and research funding becomes are established. [7]

Zheng et al. [6] introduced PharmKG, which solves the problems of sparse and noisy data sets, insufficient modeling methods, and non-uniform evaluation metrics. PharmKG is a multi-relational, attributed biomedical KG, composed of over 500 000 individual interconnections among genes, drugs, and diseases, and with 29 relation types over a vocabulary of approximately 8000 disambiguated entities. PharmKG was constructed by a novel heterogeneous graph attention neural network (HRGAT). For baselines, Zheng et al. [6] offered nine state-of-the-art (SOTA) KG embedding (KGE) methods and a new biological, graph neural network-based KGE method that uses a combination of heterogeneous domain features and global network structure. Lastly, Zheng et al. [6] observed that the HRGAT model outperforms the SOTA embedding methods. They also provided insights and guidelines on using KG in biomedicine.

For the chemical domain, risk management of hazardous chemicals is key to the safety of life and property. Given that extensive information and knowledge of hazardous chemicals is stored in isolated and various databases, Zheng et al. [8] proposed a chemical knowledge KG to fill in the information gap between decentralized databases. To extract useful information from multi-source data, they utilized the pre-trained BERT-CRF model to conduct NER for completing KG. Overall, the model achieved good results in the NER task in the chemical industry.

In the legal domain, Schneider et al. [9] proposed a legal KG (LKG), which adopts the statistical language model, BERT, BiLSTM, and CRF for the NER task. For entity linking (EL), they utilized either the BERT or DistilBERT model, which is a small, fast, affordable and light transformer model based on the BERT architecture. Overall, LKG interlinks numerous documents through linguistic and semantic information.

The summary of domain-specific KG is shown in Table II. Legal empirical research combined with artificial intelligence (AI) technology has received considerable attention in recent years. However, only a few studies have focused on the application of intelligent computing in encrypted currency money-laundering prevention case text analysis [7]. Accordingly, constructing a KG of cryptocurrency anti-money laundering is crucial in understanding the connection between crime patterns and emerging financial technology. Furthermore, analyzing and developing the relation types that need to focus on cryptocurrencies money laundering is also important to understand money laundering cases that use cryptocurrency as medium.

TABLE II. SUMMARY OF DOMAIN-SPECIFIC KNOWLEDGE GRAPH

Authors	Domains	Construction Algorithms	Evaluation Measures	Performance
Li et al. [1]	Medical	Auto TransX	N/A	N/A
Li et al. [2]	Education	BERT-BiLSTM-CRF	Precision, Accuracy, Recall, F1-score	85.32, 85.72, 85.52 (%)
Xu et al. [4]	Biomedical	BioBERT	F1-score	98.09 (%)
Zheng et al. [6]	Biomedical	HRGAT	MRR, Hits@1, Hits@3, Hits@10, Hits@100	0.154, 0.075, 0.172, 0.315, 0.649
Zheng et al. [8]	Chemical	BERT-CRF	Precision, Recall, F1-score	94.11, 94.50, 94.28 (%)
Schneider et al. [9]	Legal	BERT-BiLSTM-CRF	Accuracy, Precision, Recall, F1-score	88.82, 89.26, 84.39, 86.76 (%)

III. RESEARCH METHOD AND SYSTEM ARCHITECTURE

This study adopted the research methodology of systems development in information systems research [14]. We present the following five major stages for developing the relation types of cryptocurrency anti-money laundering KG system development methodology process.

1) Construct a conceptual framework

We will describe research problems and analyze and identify the relation types of cryptocurrency anti-money laundering KG. Thereafter, we will investigate the requirements of the cryptocurrency anti-money laundering KG system.

2) Develop a system architecture

We will develop a modular and expandable cryptocurrency anti-money laundering KG system architecture. Thereafter, we can identify the core relations of the cryptocurrency money laundering and describe the correlation among entities in the cryptocurrency anti-money laundering KG.

3) Analyze and design the system

We will design a knowledge base that can implement the functions of the cryptocurrency anti-money laundering KG system. Thereafter, we will design various workable schemes for core relation types and choose a superior scheme from them.

4) Build the prototype system

By establishing the cryptocurrency anti-money laundering KG system, we will further understand the concepts of the core and design of the relation types of cryptocurrency anti-money laundering KG.

5) Observe and evaluate the system

We will observe the use of the cryptocurrency anti-money laundering KG system through case studies or field research. Moreover, we will evaluate the effectiveness of the identified relation types of cryptocurrency anti-money laundering KG system.

This study proposed an AI architecture for cryptocurrency anti-money laundering KG, as shown in Fig. 1. The proposed system architecture comprises four principal components: knowledge acquisition, named entity recognition, identifying relation types of cryptocurrency anti-money laundering KG, and KG completion. We concentrated on NER and relation types identification. To identify relation types, Chen et al. [15] defined four relation types often observed in drug-related crimes in China to construct a legal triplet extraction system. To better understand the relationship among the numerous cryptocurrency money laundering-related entities, this study developed seven relation types concentrated solely on cryptocurrency money-laundering cases.

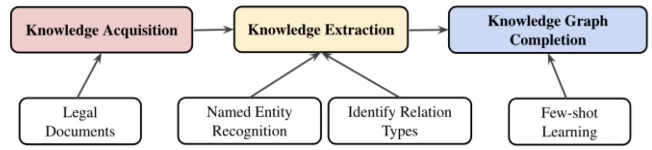

Fig. 1. Proposed architecture of AI for cryptocurrency anti-money laundering knowledge graph.

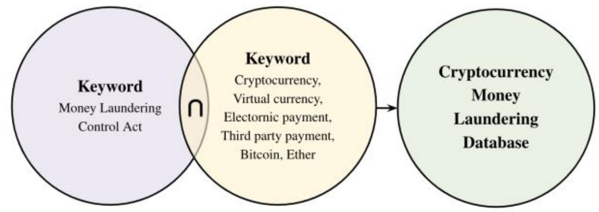

Fig. 2. Diagram of collecting judicial data.

IV. DATA ANALYSIS AND DISCUSSIONS

A. Knowledge Acquisition

Cryptocurrency-related data are collected from the legal and regulation retrieval system of the judicial court. First, this study collected data with the keyword money laundering from January 1, 2018 to December 31, 2021. From the four-year data, the cryptocurrency-related data were extracted to form a cryptocurrency money laundering knowledge base. The diagram of the judicial data collection process is shown in Fig. 2. For money laundering- and cryptocurrency-related data, the statistics is presented in Table III. As shown in Table III, the proportion of cryptocurrency-related data to money laundering-related data is approximately 10%. Given the small data set, we endeavored to utilize the few-shot learning algorithm for constructing the cryptocurrency anti-money laundering KG [7] so that the machine learning model can be trained without an extensive training data set.

B. Named Entity Recognition

This research utilized the traditional Chinese BERT-based transformers models [16] to complete the NER task. Moreover, the current study has 18 entity types. For each entity type, the description of entity types are described in Table VI, and the statistics of the NER result is listed in Table IV. Furthermore, the statistics of the NER result (in percentage) is listed in Table V. Tables IV and V show that in consolidating results for four years, the top eight high-quantity entity types in the NER task

TABLE III. STATISTICS OF THE COLLECTED DATA

Year	# money laundering-related data	# cryptocurrency-related data
2018	255	16
2019	373	11
2020	433	28
2021	3356	354
Total	4417	409

TABLE IV. STATISTICS OF THE NER RESULTS

Entity Types	# Year 2018	# Year 2019	# Year 2020	# Year 2021
PERSON	21438	27509	35087	161244
ORDINAL	13738	23371	27677	162492
CARDINAL	9878	13520	18615	95423
DATE	9122	13670	17325	93785
ORG	6814	12532	14109	81349
LAW	6069	9902	11480	69120
MONEY	4669	7541	9311	51222
GPE	1746	3879	4011	15475
TIME	1681	2264	3093	24652
PERCENT	560	305	1157	796
FAC	270	505	582	2281
LOC	133	375	280	1041
PRODUCT	125	246	519	2039
NORP	91	221	312	1552
QUANTITY	35	47	74	338
WORK_OF_ART	19	51	85	385
LANGUAGE	16	130	63	161
EVENT	12	54	81	62
Total	76416	116122	143861	763417

TABLE V. STATISTICS OF THE NER RESULTS (IN PERCENTAGE)

Entity Types	Year 2018 (%)	Year 2019 (%)	Year 2020 (%)	Year 2021 (%)
PERSON	28.05	23.69	24.39	21.12
ORDINAL	17.98	20.13	19.24	21.28
CARDINAL	12.93	11.64	12.94	12.5
DATE	11.94	11.77	12.04	12.28
ORG	8.92	10.79	9.81	10.66
LAW	7.94	8.53	7.98	9.05
MONEY	6.11	6.49	6.47	6.71
GPE	2.28	3.34	2.79	2.03
TIME	2.2	1.95	2.15	3.23
PERCENT	0.73	0.26	0.8	0.1
FAC	0.35	0.43	0.4	0.3
LOC	0.17	0.32	0.19	0.14
PRODUCT	0.16	0.21	0.36	0.27
NORP	0.12	0.19	0.22	0.2
QUANTITY	0.05	0.04	0.05	0.04
WORK_OF_ART	0.02	0.04	0.06	0.05
LANGUAGE	0.02	0.11	0.04	0.02
EVENT	0.02	0.05	0.06	0.01
Total	100	100	100	100

are PERSON, ORDINAL, CARDINAL, DATE, ORG, LAW, MONEY, and GPE (in descending order). Moreover, we surveyed the cryptocurrency money laundering-related data, which we acquired from the legal and regulation retrieval system of the judicial court in Taiwan. A closer inspection of the judicial data shows that PERSON, ORG, LAW, and MONEY are markedly and informative enetity types for this research in constructing the cryptocurrency anti-money laundering KG. In summary, what stands out in the table is the key entity type that will be focused on the cryptocurrency anti-money laundering KG.

TABLE VI. DESCRIPTION OF NER ENTITY TYPES

Entity Types	Descriptions
PERSON	People, including fictional
ORDINAL	"first," "second"
CARDINAL	Numerals that do not fall under another type
DATE	Absolute or relative dates or periods
ORG	Companies, agencies, institutions, etc.
LAW	Named documents made into laws
MONEY	Monetary values, including units
GPE	Countries, cities, states
TIME	Times smaller than a day
PERCENT	Percentage (including "%")
FAC	Buildings, airports, highways, bridges, etc.
LOC	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc. (Not services)
NORP	Nationalities or religious or political groups
QUANTITY	Measurements, as of weight or distance
WORK_OF_ART	Titles of books, songs, etc.
LANGUAGE	Any named languages
EVENT	Named hurricanes, battles, wars, sports events, etc.

TABLE VII. JUDICIAL7: IDENTIFIED RELATION TYPES OF CRYPTOCURRENCY ANTI-MONEY LAUNDERING KNOWLEDGE GRAPH

ID	Relations
#1	犯 / commits
#2	犯罪所得, 取得, 獲取 / owns
#3	提供, 交付, 交予 / provides
#4	幫助 / abets
#5	匯入 / deposits
#6	匯出 / withdraws
#7	匯款/ transfers

C. *Identified Relation Types of Anti-Money Laundering KG*

According to the NER result, which incorporates 18 entity types, and information in judicial data, our study developed the relation types of KG involving cryptocurrency anti-money laundering. The identified relation types of cryptocurrency anti-money laundering KG are called "Judicial7" and listed in Table VII. For each relation type, we extracted the example from the cryptocurrency money laundering-related judicial data, which was discussed in the knowledge acquisition section and shown in Table III. Overall, this study focused on seven relation types: commits, owns, provides, abets, deposits, withdraws, and transfers.

V. CONCLUSIONS

To construct a cryptocurrency anti-money laundering KG, we first conducted a NER task and surveyed the judicial data related to cryptocurrency money laundering. From the results of the previously discussed two stages, we indicated the legal relation types that may be focused on cryptocurrency anti-money laundering KG.

The major contribution of this research is we have proposed "Judicial7," which is the identified relation types for cryptocurrency anti-money laundering KG. In the future, this research will contribute to the application of the proposed relation types of cryptocurrency anti-money laundering KG for constructing a KG in the legal field.

ACKNOWLEDGEMENT

This research was supported in part by Ministry of Science and Technology (MOST) under the grant number 110-2410-H-305-013-MY2 and National Taipei University (NTPU) under the grant number 111-NTPU-ORDA-F-001 and 111-NTPU-ORDA-F-003.

REFERENCES

[1] Z. Li *et al.*, "DeepKG: an end-to-end deep learning-based workflow for biomedical knowledge graph extraction, optimization and applications," *Bioinformatics,* vol. 38, no. 5, pp. 1477-1479, 2021, doi: 10.1093/bioinformatics/btab767.

[2] N. Li, Q. Shen, R. Song, Y. Chi, and H. Xu, "MEduKG: A Deep-Learning-Based Approach for Multi-Modal Educational Knowledge Graph Construction," *Information,* vol. 13, no. 2, p. 91, 2022.

[3] 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,* 2021.

[4] J. Xu *et al.*, "Building a PubMed knowledge graph," *Scientific data,* vol. 7, no. 1, pp. 1-15, 2020.

[5] F. M. J. Teichmann and M.-C. Falker, "Money laundering via cryptocurrencies–potential solutions from Liechtenstein," *Journal of Money Laundering Control,* vol. 24, no. 1, pp. 91-101, 2021.

[6] S. Zheng *et al.*, "PharmKG: a dedicated knowledge graph benchmark for bomedical data mining," *Briefings in Bioinformatics,* vol. 22, no. 4, 2020, doi: 10.1093/bib/bbaa344.

[7] M.-Y. Day, "Artificial intelligence for knowledge graphs of cryptocurrency anti-money laundering in fintech," presented at the Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Virtual Event, Netherlands, 2021.

[8] X. Zheng, B. Wang, Y. Zhao, S. Mao, and Y. Tang, "A knowledge graph method for hazardous chemical management: Ontology design and entity identification," *Neurocomputing,* vol. 430, pp. 104-111, 2021.

[9] J. M. Schneider *et al.*, "Lynx: A knowledge-based AI service platform for content processing, enrichment and analysis for the legal domain," *Information Systems,* vol. 106, p. 101966, 2022.

[10] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, and Z. Duan, "Knowledge graph completion: A review," *Ieee Access,* vol. 8, pp. 192435-192456, 2020.

[11] B. Abu-Salih, "Domain-specific knowledge graphs: A survey," *Journal of Network and Computer Applications,* vol. 185, p. 103076, 2021.

[12] X. Chen, S. Jia, and Y. Xiang, "A review: Knowledge reasoning over knowledge graph," *Expert Systems with Applications,* vol. 141, p. 112948, 2020.

[13] Y. Wang, H. Zhang, G. Shi, Z. Liu, and Q. Zhou, "A model of text-enhanced knowledge graph representation learning with mutual attention," *IEEE Access,* vol. 8, pp. 52895-52905, 2020.

[14] J. F. Nunamaker Jr, M. Chen, and T. D. Purdin, "Systems development in information systems research," *Journal of management information systems,* vol. 7, no. 3, pp. 89-106, 1990.

[15] Y. Chen, Y. Sun, Z. Yang, and H. Lin, "Joint entity and relation extraction for legal documents with legal feature enhancement," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1561-1571.

[16] P.-H. Li, T.-J. Fu, and W.-Y. Ma, "Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 05, pp. 8236-8244.