Dynamic Knowledge Graph Based Construction of Quality Infrastructure System for Non-API Oil Country Tubular Goods

Xiangpeng Chen¹, Juntai Xie¹*, Jianmin Gao¹, Rongxi Wang¹ and Jiandong Jiang²

¹ State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China
² Key Laboratory of Quality Infrastructure Efficacy Research, State Administration for Market Regulation, Chaoyang District, Beijing, 100028, China
Email: xiejuntai@xjtu.edu.cn

Abstract. With the deep exploitation of oil and gas resources, the non-API oil country tubular goods (OCTG) adapted to specific environments are used widely. Therefore, how to effectively characterize the quality connotation of non-API OCTG to ensure their quality has become a challenge for the petroleum industry. We propose a dynamic knowledge graph of Quality Infrastructure (QI) to solve the problems of the diversity of non-API OCTG quality influencing factors, the concealment of the relationship, and the ambiguity of the mechanism of quality improvement. Firstly, a knowledge graph ontology framework of quality infrastructure is constructed, which realizes the effective combination of product characteristics and quality basic elements. Secondly, based on the professional dictionary in the field of OCTG, entity recognition adopts the entity recognition method of LDA-BiLSTM-CRF, which effectively improves the recognition accuracy of professional vocabulary. Finally, the relationship between entity types is defined as the edge of the knowledge graph; the graph embedding method is used to supplement the edge connection and calculation weight of the knowledge graph. The QI knowledge graph constructed with this technology can well describe the quality connotation of non-API OCTG, and provide opinions and methods for guaranteeing and improving the quality of non-API OCTG.

Keywords. Knowledge Graph, Ontology Framework, LDA, Graph Embedding, Non-API OCTG

1. Introduction
In order to meet the requirements for safe and stable exploitation of oil and gas resources in ultra-deep wells, high temperature and pressure and different corrosive environments in oil and gas exploration and development, the demand for high performance OCTG represented by non-API OCTG is increasing. The non-API OCTG is a personalized product developed by the manufacturer in order to meet certain special performance requirements according to user needs to adapt to the performance requirements of different environments. It is a patented product of each manufacturer [1, 2]. Insufficient quality of non-API OCTG will cause huge social problems, leading to a series of safety accidents such as personal injury, environmental pollution, and equipment damage during the operation of oil and gas wells [3]. The quality of non-API OCTG will be affected by a series of factors [4], including the physical and chemical characteristics, the service environment, and the level of the...
QI elements in the whole life cycle. It is precisely because of the diversity of these quality influencing factors, the secret and complexity of the coupling relationship, and the ambiguity of the mechanism of quality improvement. It is difficult for the petroleum industry and related companies to uniformly manage the information, and provide advice on quality control activities during the entire life cycle of non-API OCTG.

As an advanced and effective knowledge management method, knowledge graph is developed on the basis of semantic network. Knowledge graph related technologies include knowledge extraction [5, 6], knowledge representation [7, 8], knowledge fusion [9, 10], etc., which are widely used in intelligent search [11], in-depth question and answer [12], social networks [13] and vertical industries [14]. As a knowledge organization and retrieval technology in the era of big data, knowledge graph provides a more effective way for the expression, organization, management and utilization of massive, heterogeneous and dynamic big data, making the network more intelligent, which is closer to human cognitive thinking [15].

Knowledge graphs are gradually developing in the direction of professional fields, including medicine [16, 17], chemical industry [18], e-commerce [19], etc. These studies have achieved beneficial results, strengthened people's understanding of professional fields, and promoted the development of related industries. Different from the establishment of a knowledge graph for non-API OCTG products, a knowledge graph of the quality for OCTG products can be constructed. In addition to OCTG itself, the influence of its QI elements is emphasized. We need to consider the following points: (1) The service environment of the OCTG is complicated, which has an important impact on the quality of the product; (2) The physical and chemical properties of the non-API OCTG standard are very different from the API standard products. It is necessary to analyse the important physical and chemical properties; (3) The OCTG usually have different quality requirements under different service environments and different physical and chemical properties; (4) The focus of the research is on the quality of non-API OCTG products to characterize the connotation. It is necessary to integrate the basic elements of QI into the non-API OCTG products.

The main innovations of this article include:

1) The knowledge graph is applied to the petroleum field for the first time, with emphasis on characterizing the quality connotation of non-API OCTG. By analysing the characteristics of non-API OCTG, a hierarchical structure of QI is constructed, which provides a basis for guaranteeing and improving the quality of non-API petroleum special pipes;

2) A “LDA-BiLSTM-CRF” entity recognition technology is proposed. For unstructured and semi-structured speech data, first use LDA for topic analysis, and then use BiLSTM-CRF for entity recognition, which improves the accuracy of entity recognition. And LDA can discover new entity categories and realize the dynamic expansion of the knowledge graph;

3) The relationship completion technology of graph embedding is used to convert the knowledge graph into a low-dimensional dense vector. On this basis, data analysis is performed on entities of different categories, and the hidden relationships between categories and the strengths of relationships are explored.

2. Ontology Construction

The construction of domain ontology. This research adopts the seven-step method of constructing ontology of Stanford University [20], and adopts a combination of top-down and bottom-up methods in the definition of classes and class hierarchy. An important purpose of this research is to establish a QI knowledge graph of the OCTG. First, the top-down construction method is adopted to define the classes of the knowledge graph from an overall perspective; the classes and attributes to be constructed are considered. When it meets and covers the actual data, a bottom-up approach is used to complete the definition of categories and attributes. The combination of the two can build a complete and reasonable QI knowledge graph of non-API petroleum special pipes.
2.1. Determine the Field and Category of Knowledge Graph
The object of this research is mainly non-API OCTG. The non-API OCTG can be said to be a special kind of API OCTG with higher quality requirements. Therefore, the domain of the knowledge graph is extended to the field of petroleum special pipes, including API standard OCTG and non-API OCTG. At the same time, the OCTG products are mainly used in petroleum engineering (the petroleum engineering in this study mainly includes: well construction engineering (drilling and completion), testing engineering (logging, well testing, oil testing and production), oil production (oil production). In order to obtain richer data, stronger applicability of the established knowledge graph and reduce the difficulty of data screening in the process of establishing the knowledge graph. We set the extension of the knowledge graph field to be in petroleum engineering, and specifically focus on the knowledge of OCTG products.

2.2. Determine the Class Hierarchy
The goal of this research is to build a QI knowledge graph, which need to solve the problem of how to combine OCTG and QI. Through research, finally determined the class hierarchical structure of the knowledge graph in 2 categories and 8 sub-categories. It includes four sub-categories of the "product nature" category: service environment, material chemistry, structural mechanics and quality requirements; 8 sub-categories of the "quality foundation" category: metrology, standards, inspection and certification. There is a progressive hierarchical relationship between the two categories. "product nature" is the bottom level; "quality foundation" is the upper level, reflecting the idea from the bottom product to the upper quality.

2.2.1. Determination of the Hierarchical Structure of Product Quality Characteristics. The study divided the "product properties" into four categories: service environment, material chemistry, structural mechanics, and quality requirements. The four sub-categories can be further divided into 8 sub-categories of geological conditions, working environment, materials, chemical composition, structure, mechanical properties, service quality, and other requirements. At the same time, each sub-category can be further subdivided according to the situation, such as geological conditions can be subdivided into temperature conditions and pressure conditions. As shown in figure 1 below:

![Figure 1. Hierarchical structure of product quality characteristics.](image)

2.2.2. Determination of the Hierarchical Structure for QI System. The QI system is divided into four sub-categories: metrology, standards, inspection and testing, and certification. The four sub-categories can be further divided into metrology reference level, metrology instrument condition, metrology value traceability, metrology personnel, standard system level, standard implementation level,
standard-related personnel, standard-related systems, inspection and testing implementation level, inspection and testing equipment, inspection and testing system, inspection and testing personnel, certification policy orientation, certification capabilities, certification implementation level, and certification system level. These 16 subcategories are as shown in figure 2.

![Figure 2. Hierarchical structure of quality infrastructure system.](image)

### 3. Entity and Relationship Mining

Since the petroleum industry involves national security, and the constructed QI knowledge graph involves non-API OCTG full life cycle data, there are relatively few structured data. Mainly extract entities from unstructured and semi-structured data to enrich the knowledge graph. The content of this part mainly includes domain dictionary construction and entity extraction technology based on LDA-BiLSTM-CRF.

Construction of Dictionary in the Field of Petroleum Engineering: At present, most of the entity libraries of the knowledge graph are entity libraries in general fields, and the number of entity libraries in special fields is relatively small. And there is no authoritative entity database in the field of petroleum engineering, especially in the field of petroleum special pipes. In order to improve the accuracy of entity extraction and relationship extraction in the subsequent automatic knowledge graph construction process. In this section, an entity database in the field of petroleum engineering containing more than 10,000 vocabularies are established by crawler technology and artificially, such as tube body, Oil well pipeline, magnetic flux leakage testing, Drilling fluid annulus flow pattern, etc.

Named Entity Recognition (NER) is to extract entities (words with specific meaning or strong reference) from unstructured input text, such as rounded corners and bosses under the physical structure category of petroleum pipe products. Named entity recognition technology is an indispensable part of various natural language processing technologies such as information extraction, information retrieval, knowledge graphs, machine translation, and question answering systems.

Because the knowledge graph in this research contains many categories-there are 24 categories in total. The quality of unstructured data varies, and the required knowledge is fragmented. If entity recognition is performed directly on such data, the effect is not very good. To this end, this paper proposes a method of LDA-BiLSTM-CRF. First, extract the data of each type from the text data through the topic model (LDA) method, and determine which of the 24 categories it belongs to. (Note: When it does not belong to any of the 24 categories, it is judged by the key topic words extracted from
it. When these topic words are more related to OCTG, they are defined as a new category. Otherwise, they are regarded as Invalid data is discarded); the method of BiLSTM-CRF is used to learn each type of data separately and identify the entities in it.

Since this research has constructed the QI knowledge graph, there is no need to clarify the specific relationships between entities. The focus of the research is whether there is a relationship between entities and the strength of the relationship. Therefore, 16 kinds of relationships among 24 categories are directly defined as relationships between entities, as shown in the following table 1.

| No. | Entity class          | Entity subclass                  | Relationship symbol |
|-----|-----------------------|----------------------------------|---------------------|
| 1   | materials             | standard system level            | Ma-st               |
| 2   | materials             | inspection and testing equipment | Ma-it               |
| 3   | chemical composition  | standard system level            | Ch-sl               |
| 4   | chemical composition  | inspection and testing equipment | Ch-it               |
| 5   | structure             | standard system level            | St-sl               |
| 6   | structure             | inspection and testing equipment | St-it               |
| 7   | mechanical properties | standard system level            | Mp-sl               |
| 8   | mechanical properties | inspection and testing equipment | Mp-it               |
| 9   | standard system level | inspection and testing system    | Sl-it               |
| 10  | standard system level | inspection and testing equipment | Sl-it               |
| 11  | inspection and testing equipment | metrology reference level | It-ml              |
| 12  | inspection and testing equipment | metrology instrument condition | It-mi              |
| 13  | inspection and testing system | certification system level     | It-cl               |
| 14  | product               | certification system level       | Pr-cl               |
| 15  | product A1            | product A2                       | si-re               |
| 16  | product A             | product B                        | co-re               |

After completing the construction of the QI knowledge graph, through the graph embedding algorithm DeepWalk, the graph structure information of the constructed knowledge graph is transformed into a low-dimensional dense vector, and the strength of the coupling relationship between the categories is calculated. Identify the hidden relationship between the classes, and then draw the key factors that affect the quality of OCTG, and on this basis, carry out the quality assurance and improvement of OCTG.

4. Experimental Verification
According to the proposed method, a QI knowledge graph for non-API OCTG is actually constructed. Verify the validity of the QI knowledge graph proposed in this article.

4.1. Data Introduction
The data collected by the research mainly includes three categories:
- Professional and authoritative books in the field of petroleum special pipes (electronic version);
- Related electronic literature in the field of petroleum special pipes (crawler acquisition);
- Information provided by relevant organizations of petroleum special pipes (electronic Version).

The specific form and use of the relevant data are as table 2:
Table 2. Data situation.

| No. | Data Source | Size  | Type          | Quality | Use              |
|-----|-------------|-------|---------------|---------|------------------|
| 1   | Encyclopedia of China Petroleum Exploration and Development" Abstracts of Petroleum Pipes and Instruments | The internet | 50M | Unstructured data | low | Dictionary construction |
| 4   | The internet | 1352KB | Unstructured data | middle | Dictionary construction |
| 5   | The internet | 270KB | Unstructured data | middle | Dictionary construction |
| 6   | The internet | 39KB | Unstructured data | middle | Dictionary construction |
| 7   | OCTG Organization | 3GB | Unstructured data | middle | Knowledge Graph |

4.2. Entity Recognition

4.2.1. LDA Theme Analysis. In this study, the failure analysis report of petroleum pipes was used as data, and three types of unstructured data texts including chemical composition, metallographic structure and mechanical properties were selected for LDA theme analysis. The accuracy and data volume of text data subject recognition are shown in the figure 3 below:

![Figure 3. LDA model accuracy and data volume.](image)

4.2.2. BiLSTM-CRF. First of all, the BIO labeling method is used to label the text data. Second, based on the three types of text data obtained by using the LDA topic model, the YEDDA open source tool is used to label each type of text data separately. And a total of eight types of entities are labeled. Finally, when actually training the model, select 70% as the training set, 20% as the test set, and 10% as the verification set. The following is an example of the chemical composition classification under the LDA theme model. Four classification tasks were carried out under the chemical composition classification, including "standard", "testing Equipment", "chemical organization", and "product". After the end of the labeling work, we encode the characters with the Bert model file Chinese_l-12_h-768_a-12 trained by Google. And the accuracy curve obtained by training is shown in the figure 4.
4.3. Hidden Relationship Mining and Weight Determination

We enter the fragment of the knowledge graph shown in figure 5 into DeepWalk [21] algorithm (different colours represent different categories), and use '--representation-size' to be set to 64; '--walk-length' to be set to 40; '--number- Walks' is set to 10. Finally, the coupling relationship and weight between the third types of nodes are obtained as shown in the following table 3.

| Table 3, Calculation results of coupling relationship. |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
| Roundness | Cr content | Small grain size | Anti-crushing | S corrosion resistance |
| Roundness | 0 | 557.1 | 297.4 | 295.9 | 659.4 |
| Cr content | 557.1 | 0 | 52.3 | 677.6 | 7.7 |
| Small grain size | 297.4 | 52.3 | 0 | 403.9 | 96.7 |
| Anti-crushing | 295.9 | 677.6 | 403.9 | 0 | 796.5 |
| S corrosion resistance | 659.4 | 7.7 | 96.7 | 796.5 | 0 |

It can be found that ‘Cr content ’and ‘S corrosion resistance’ are very closely related. According to the actual situation, there is a close correlation between Cr content in the material and hydrogen sulfide corrosion protection, which is in line with the actual situation. When there is a requirement for S corrosion resistance for OCTG, the metallographic structure and grain size of OCTG materials are also required to be small. These show that the DeepWalk algorithm can well discover the hidden relationships in the knowledge graph and judge the strength of the relationship.

Figure 4. Train accuracy curve.

Figure 5. Fragment of knowledge graph.
5. Conclusion
The research in this paper mainly clarifies the quality influencing mechanism of non-API OCTG and characterizes the quality connotation of non-API OCTG from a large amount of scattered data. This paper designs the construction process of the non-API OCTG QI knowledge graph. To explorative, the ontology framework of the knowledge graph of "product nature-quality basis" is proposed, and quality is integrated into the representation of the knowledge graph. At the same time, an entity recognition method of LDA + BiLSTM + CRF is proposed. Divide the data into segments, and perform entity recognition based on the subject of each segment of data. To a certain extent, it overcomes the impact of low data quality and effectively improves the quality of entity recognition. The proposed QI knowledge graph aggregates scattered OCTG quality knowledge and integrates and supplements expert experience. It has greatly deepened the understanding of the connotation of OCTG quality and laid the foundation for the quality assurance and improvement of non-API OCTG products.

Acknowledgments
This research was supported by the National Key R&D Program of China (No. 2019YFF0217500) and National Key Laboratory of Quality Infrastructure Effectiveness Research Open Research Fund under Grant No. KF20180301. The authors would also like to thank the Institute of New Energy Equipment and Quality Engineering of Xi’an Jiaotong University, for their support and advice.

References
[1] Gao X, et al. 2020 CO2 corrosion behavior of high-strength and toughness V140 steel for oil country tubular goods Journal of Materials Engineering and Performance 29 8451–8460.
[2] Liu Z, et al. 2017 The domestic and foreign research status of high strength and high toughness of OCTG Petroleum Tubular Goods & Instruments 5-8.
[3] Yo S, et al. 2018 Studies on high temperature-durable IC chip label for FRP OCTG Gpi Journal 1-4.
[4] Saad, et al. 2014 OCTG premature failures due to metallurgical variance and differential electrochemical behaviour of exterior and interior walls International Journal of Oil Gas & Coal Engineering 2(3) 28-35.
[5] Deng K, Zhang X and Liu J 2018 Distant supervision for relation extraction via retrieval-based neural networks 2018 IEEE 4th International Conference on Computer and Communications (ICCC) p 1753-1762.
[6] Ji G L, et al. 2017 Distant supervision for relation extraction with sentence-level attention and entity descriptions Thirty-First Aaai Conference on Artificial Intelligence p 3060-3066.
[7] Xie R B, et al. 2016 Representation learning of knowledge graphs with entity descriptions Thirtieth Aaai Conference on Artificial Intelligence p 2659-2665.
[8] Niu Y L, et al. 2017 Improved word representation learning with sememes Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017) vol 1 p 2049-2058.
[9] Dong X L, et al. 2014 Knowledge vault: a web-scale approach to probabilistic knowledge fusion Proceedings of the 20th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining (KDD’14) p 601-610.
[10] Preece A, et al. 2000 The KRAFT architecture for knowledge fusion and transformation Knowledge-Based Systems 13(2-3) 113-120.
[11] Zhu Y, et al. 2017 Intelligent learning for knowledge graph towards geological data Scientific Programming 2017(pt.1) 1-13.
[12] Yang Z, et al. 2021 Design and research of intelligent question-answering (Q&A) system based on high school course knowledge graph Mobile Networks and Applications 1-7.
[13] Elezaj O, et al. 2019 Towards designing a knowledge graph-based framework for investigating and preventing crime on online social networks International Conference on e-Democracy p 181-194.
[14] Zhao X, et al. 2020 Application of neural network based knowledge graph in vertical industry
Journal of Physics Conference Series 1584 012018.

[15] Mccusker J P, et al. What is a Knowledge Graph? p 2-14.

[16] Zheng Z Q, et al. 2020 TCMKG: a deep learning based traditional chinese medicine knowledge
graph platform 11th IEEE International Conference On Knowledge Graph (ICKG 2020) p 560-564.

[17] Yu T, et al. 2017 Knowledge graph for TCM health preservation: Design, construction, and
applications Artificial Intelligence in Medicine 77 48-52.

[18] Zheng X, et al. 2020 A knowledge graph method for hazardous chemical management: ontology
design and entity identification Neurocomputing 430 (21) 104-111.

[19] Li F L, et al. 2020 AliMe KG: Domain Knowledge Graph Construction and Application in E-
Commerce ACM p 2581-2588.

[20] Jones D, Benchcapon T and Visser P 1998 Methodologies for Ontology Development Capon pp
62-75.

[21] Perozzi B, et al. 2014 DeepWalk: online learning of social representations Proceedings of the
20th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining (KDD’14) p 701-710.