GWKG: semi-automatic construction of the Great Wall Knowledge Graph

1st Yizhan Feng
North China Electric Power University
Beijing, China
yizhan.feng@ncepu.edu.cn

2nd Jing Teng
North China Electric Power University
Beijing, China
jing.teng@ncepu.edu.cn

3rd Honglei Zhang
North China Electric Power University
Beijing, China
honglei.zhang@ncepu.edu.cn

2021 China Automation Congress (CAC) | 978-1-6654-2647-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAC53003.2021.9727293

***Abstract*—This study proposes to construct the Great Wall knowledge graph (GWKG) in a semi-automatic way. First, under the guidance of domain experts, we build the professional dictionary and ontology layer, where BERT is applied for Named Entity Recognition. The resulting entities are clustered by Word2Vec to automatically refine the ontology layer. Then, we conduct Relation Extraction based on semi-supervision, link entities to encyclopedia websites, and obtain semi-structured information by crawler technology for attribute filling. Finally, we visualize the GWKG and report the results of multiple data formats. The GWKG consists of 34 ontology concepts, 33 types of relations, more than 6,000 entities and attributes, and 720,000 Chinese corpora.**

***Index Terms*—knowledge graph, the Great Wall, ontology, semi-automatic**

I. Introduction

Google first proposed Knowledge Graph (KG) in 2012. Since then, KGs are widely used in semantic search, intelligent question answering, and recommendation systems. KG is a multi-relational directed graph composed of a massive amount of nodes and edges in the form of triples <Subject, Predicate, Object>. The nodes can be either entities or concepts, while the edges can describe relations and attributes. KGs describe and store the entities, concepts, and relations among objective facts in a highly structured form. Entities represent specific substances in the objective world, while concepts are abstract representations of objective substances. For example, the Badaling Great Wall is a scenic spot, where the Badaling Great Wall is an entity, and the scenic spot is a concept. Attributes describe certain aspects of concepts or entities, such as the level of scenic spots. Relations can be regarded as a particular type of attribute. When an attribute value of an entity is also an entity, meaning that this attribute is essentially a relation.

KGs can be divided into open KGs and vertical KGs according to the size of the application domain. The open KGs are general and knowledgeable, such as Freebase, YAGO, NELL, DBpedia, and Wikidata, which have proven their advantages in supporting many applications. However, they lack in-depth information in specific fields, leading to the difficulty of applying domain knowledge. The vertical KGs are oriented to specific fields, such as finance, e-commerce, medical treatment, education, which often have more sources of domain knowledge, more complex structure, and higher quality than the open KGs. KG of the Great Wall cultural heritage belongs to the vertical field.

As one of the most prominent linear cultural heritages globally, the Great Wall has outstanding value worldwide. The length of construction time, the wide distribution area, and the significant influence of the Great Wall are unmatched by similar cultural heritages. The Great Wall stands in the land of China with its majestic momentum of over two thousand years up and down, more than 100,000 miles in vertical and horizontal directions. In 2017, the "Great Wall Cultural Belt Construction" became one of the critical contents of the Beijing City Master Plan. In 2003, United Nations Educational, Scientific and Cultural Organization (UNESCO) drafted a digital cultural heritage protection program, which is favored by the world, providing new ideas for the "live" presentation and digital inheritance of cultural heritage.

The current fully automatic construction method based on deep learning is unsuitable for constructing the GWKG, which may lead to fuzzy concepts, poor knowledge quality, and low data utilization. Therefore, the KG obtained by the fully automatic construction method cannot be connected to the field application. However, manual construction directly by domain experts is pretty time-consuming and labor-intensive. This approach cannot be extended to a large number of concepts and relations. Considering these issues, we believe that the semi-automatic construction of GWKG combined with manual annotation and deep learning is the optimal solution. The main contributions of GWKG are as follows:

- GWKG is currently the first attempt to apply KGs to the protection of the Great Wall heritage. The data comes from high-quality research texts collected by China National Knowledge Infrastructure (CNKI) and the authoritative literature database of the cultural industry.
- The construction of the ontology layer utilizes a combination of top-down and bottom-up methods to model domain concepts. Under the guidance of experts, we sort out the knowledge system of the Great Wall cultural heritage and establish a coarse-grained ontology layer from top to bottom. After Named Entity Recognition, the entities are clustered to find and make up for the blind spots of experts in combing concepts. After manual verification, we build a fine-grained

This work was supported by the National Natural Science Foundation of China under Grant No. 61503137 and No. 61871181, and the Fundamental Research Funds for the Central Universities under Grant No. 2020MS017.

ontology layer from the bottom up. Finally, a mature and complete ontology layer is constructed.

- The whole process of GWKG is based on semi-automatic construction, using semi-supervised and unsupervised deep learning methods. With the support of the domain dictionary, we adopt the natural language processing tool Jiagu combined with the pre-trained model Bidirectional Encoder Representations from Transformers (BERT) [1] for Named Entity Recognition. After that, we train Word2Vec [2] on the Chinese Wiki Corpus and apply the K-means algorithm to cluster the resulting entities. Finally, Relation Extraction chooses semi-supervised RSN [3] to learn the similarity of relations from a small amount of labeled text and migrate to many unlabeled texts for semi-development clustering.
- The existing vertical domain KGs suffer from a lack of visualization. We explored visualization methods including Neo4j, Circos [4], Gephi, and D3 to promote the GWKG visualization.

II. Related Work

The goal of the KG is to describe the various entities and concepts that exist in the real world and the relations between these entities and concepts. The vertical KGs require high accuracy and depth of domain knowledge to assist various complex analysis applications or data support. Moreover, a strict and rich ontology layer should be integrating into the vertical KG. A complete ontology layer is the most prominent feature that distinguishes the vertical KGs from the open KGs. Ontology describes the data model of the KG, emphasizing concepts and the relations between them. These concepts and relations are commonly recognized, including concepts and their synonymous titles, the upper and lower relations between concepts, the attribute relations that the concepts have, the domain and value range of attributes, as well as the hypotheses and constraints on these contents.

Vertical KGs are widely applied in medical, biology, education, geography, and other industries. For example, PharmKG [5] and DrugBank [6] are involved in all aspects of the biomedical industry, including information on drugs, diseases, proteins, genes, and drug molecules. Some related research has also started in the medical field. Knowlife [7] is a medical KG constructed by automatic algorithms, which extracts and integrates data from scientific publications, encyclopedia-style healthcare portals, and online communities. In the biological field, Ozymandias [8] is a biodiversity KG constructed for the Australian fauna. The core of the ontology layer is taxa, taxonomic names, publications, journals, and people. In the education field, Gregory et al. developed HPKMT [9], which enables learners to express their understanding of the courses they have learned by independently creating concepts and defining the relation between the concepts. By leveraging heterogeneous data from the education domain, KnowEdu [10] adopts the neural sequence labeling algorithm to extract instructional concepts and employs probabilistic association rule mining on learning assessment data to identify the relations with educational significance. In terms of geography, GeoLink [11] is a KG developed for some of the largest geoscience data repositories in the United States. It includes port visits such as marine expeditions, physical sample metadata, research project funding and staffing, and authorship of technical reports. GeoLink contains millions of triples at the beginning of its creation, and it has constantly been growing. GeoNames [12] is a classic KG in the geographic field, containing more than ten million geographic information, such as region name, location, linked to specific maps.

However, compared with the construction of KGs in other fields, cultural heritage KGs are still rare. In particular, KGs related to the Great Wall are unprecedented. The British Museum Knowledge Graph [13] is the first case in the field of cultural heritage. Its ontology layer covers various metadata about museum artifacts, including historical context, associations with geographical locations, creators, discoverers, and past owners. It runs this metadata in collaborative work by creating annotations, narratives involving semantic references, and argumentations exploiting knowledge graphs as evidence. Another well-known cultural heritage knowledge graph is ArCo [14], constructed by Valentina et al. ArCo is based on the official General Catalogue of the Italian Ministry of Cultural Heritage and Activities. Its associated encoding regulations collect and validate the catalog records of all Italian Cultural Heritage properties. ArCo has fully demonstrated the potential of the cultural heritage knowledge graph, including tourism, teaching, and management. This series of successful cases inspired us to organize and construct a vertical knowledge map of the Great Wall, China's most famous world cultural heritage.

III. Methods

The logical structure of GWKG is roughly divided into the ontology layer and the data layer. The ontology layer is above the data layer and is the logical core of GWKG. Under the guidance of domain experts, the construction of the ontology layer combines the top-down and bottom-up methods to analyze and refine the concepts, relations, and attributes of the Great Wall cultural heritage. Then, a structural concept graph with various granularities is formed, which serves as the structural framework for constructing the data layer of GWKG. In the data layer, knowledge is stored in the graph database as the fundamental expression of facts in <head entity, relation, tail entity> or <entity, attribute, value pair> triples. As shown in Fig 1, the construction of GWKG is based on the original data, adopting a series of semi-automatic technical means to extract knowledge elements from the original data and store them in the data layer under the specification of the ontology layer. The construction of GWKG could be summarized into four modules: Ontology Layer Construction, Named Entity Recognition, Relation extraction, and Attribute Filling. This section will elaborate on the four modules accordingly.

A. Ontology Layer Construction

GWKG has rigorous requirements on the depth, quality, and granularity of knowledge. Therefore, a complete ontology

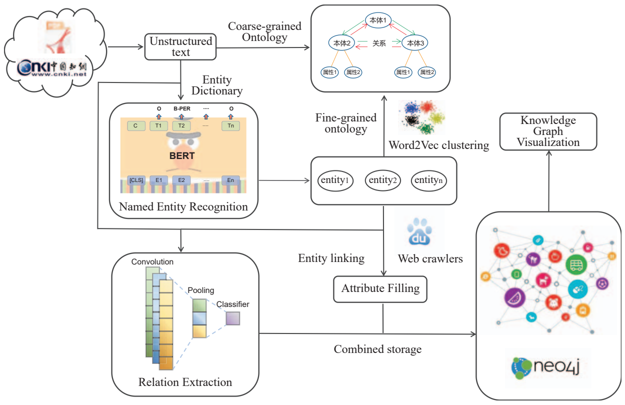

Fig. 1. The technical architecture of GWKG.

layer data model should be defined. We combine the top-down and bottom-up methods to build the ontology layer with concepts of various granularities.

Under the guidance of domain experts, we clarify the target object and determine the field and scope of GWKG. The high-quality corpus texts come from CNKI and the literature library of the cultural industry. Then we define the attributes and related data of scenic spots, buildings, people, etc.. The ontology layer of GWKG combines the characteristics of the Great Wall to define 17 concepts and their attributes, as well as 23 semantic relations. Detailed information on the ontology layer can be found in Table 1. So far, the prior data model of GWKG has been successfully constructed, named the coarse-grained ontology layer. The coarse-grained ontology layer significantly assists the subsequent construction of GWKG.

However, it is undeniable that manual construction methods are prone to errors. Even domain experts inevitably have blind spots or mistakes when combing concepts. To find and make up for experts' prior blind spots, GWKG performs Named Entity Recognition on the original corpus and then utilizes Word2Vec [2] clustering to summarize more fine-grained data patterns from the resulting entity set. Word2Vec is a word embedding training tool released by Google in 2013, which automatically learns the appearance of words in a corpus by a neural network. Word2Vec embeds words into a high-dimensional space and expresses them in the form of word vectors. The semantic relationship of word embeddings generated by Word2Vec is better reflected in the high-dimensional space, meaning that words with similar semantics are closer in the high-dimensional space. GWKG trains Word2Vec on the original text related to the Great Wall and the Chinese corpus of Wikipedia to obtain the word vector of each entity after named entity recognition. Then, GWKG clusters the entities based on the Euclidean distances between them by the K-means algorithm. After multiple pre-defined clusters K and iterative clustering, the fine-grained ontology is summarized under the evaluation and review of domain experts. The coarse-grained ontology layer is thus refined and fused to form the final complete ontology layer, with a total of 34 fine-grained concepts and 30 upper-lower relations. Table 1 lists the detailed ontology information.

B. Named Entity Recognition

Named Entity Recognition (NER) refers to recognizing entities with specific meanings in the text, including names of persons, places, organizations, and proper nouns. Because of the characteristics of NER tasks in the Great Wall field, our technical selection is a semi-automatic "Entity Dictionary + BERT" solution. On the one hand, dictionary matching can effectively improve the accuracy and speed of entity recognition. On the other hand, most current emerging NER technologies can only deal with established entities in the open field. It is challenging to deal with the specific types of entities in the vertical field. After the entity dictionary is matched, the result is guaranteed to be domain-adapted. However, there is currently no professional dictionary of Great Wall cultural heritage available to researchers. Here we build a core entity dictionary from scratch by manually labeling entities, online websites related to Great Wall, and professional terms summarized by field experts.

With the subsequent expansion of the scale of the graph, the Out Of Vocabulary (OOV) problem would inevitably arise with only the physical dictionary. The method based on deep learning has the particular ability of generalization, which can be an effective supplement to the entity dictionary. BERT [1] is a general Natural Language Processing model disclosed by Google in 2018. Since BERT was released, it attracted widespread attention from academia and industry. BERT applies the encoder connected by multi-head self-attention in the Transformer [15] to achieve greater parallelization and reduce training time. The main contribution of BERT is the idea of "pre-training + fine-tuning" to solve various problems in Natural Language Processing. Taking Chinese NER as an example, the pre-training process learns network parameters on many general corpora, including Wikipedia and Book Corpus. These corpora contain a large amount of natural text and can provide a wealth of language-related phenomena. The network parameters of GWKG are thus fine-tuned by the annotation data of the Chinese dataset MSRA-NER. The GWKG is sustainably maintained without the need to retrain a task-specific network for NER. Entity dictionary matching and BERT model predictions complement each other and can be described as the best match at the current stage.

C. Relation Extraction

Relation Extraction (RE) aims to identify the relation between entities or concepts of corpus text. Traditional RE methods mainly focus on classifying relation facts into pre-defined relation types. These methods usually apply a large amount of supervised data to learn neural sentence encoders for distributed representations. However, if we want to adopt traditional supervised learning RE methods in GWKG, a large corpus should be manually labeled, which is tedious and impractical. Furthermore, GWKG should be constantly

TABLE I

DETAILS OF ALL CONCEPTS IN THE ONTOLOGY LAYER

Concept	*Entity example*	*Attribute example*
Building	the Beacon Tower; Parapet wall; Battery	Aliases; Brief History
Scenic spot	Badaling Great Wall; MUTIANYU GREAT WALL	Attraction level; Location; Management unit
Books: Papers	The study of Jizhen Great Wall in the Ming Dynasty	Author; Creation time
Books: Historical works	Si Zhen San Guan Zhi; Jiu Bian Tu Shuo	Author; Introduction
Person: Celebrity	Xu Da; Chang Yuchun; Zhang Juzheng	Alias; Survival time; Place of birth; Achievement
Person: Chief Military Officer of Jizhen	Qi Jiguang; Zhang Xin; Chen Jingxian	The dynasty that served as the chief officer
Person: Emperor	Zhu Yuanzhang; Zhu Di; Zhu Zhanji	Alias; Brief introduction to ruling
Official: Main official position	Duke; Shangshu; Commander	Initial dynasty; Introduction
Official: Jizhen Core Official Position	General officer, General, Guerrilla	Jurisdiction and Authority; Number of troops
Official: The Guardian of Jizhen	General Tongzhou and General Xifengkou	Station; Introduction
Official: The Guerrilla of Jizhen	Miyun Zuoying guerrilla; Santun Youying guerrilla	Exclusive support; Introduction
Official: The Defense of Jizhen	Zunhua fielding; Sanhe fielding	Resident; Initial time
Official: The Tidiao of Jizhen	Taolinkou Tidiao; Damaoshan Tidiao	Aliases; Brief history
Time: Dynasty	Qin Dynasty; Yuan Dynasty; Ming Dynasty	Start time; End time
Time: Reign Title	Hongwu; Yongle; Jiajing	Starting and ending time; The reigning emperor
Vegetation: Wild plants	Castanea mollissima; Lespedeza bicolor Turcz	Common name; Plant resource; Family
Vegetation: Protected plant	Tilia amurensis Rupr.; Juglans mandshurica	Common name; Plant resource; Family
Nine Sides	Ji zhen; Chang zhen; Xuanfu zhen	Length of the Great Wall; Number of troops
Sanitation	Yongping Sanitation, LuLong Sanitation, Zhuolu Sanitation	Governance; Affiliation; Number of farms
Jizhen Military Road: General Road	Songpeng Road; Malan Road; Caojia Road	Length; Density; Slope of earth and rock
Jizhen Military Road: Military Path	Yongping Path, Jizhou Path, Miyun Path	Leading institution; Direct commanding officer
Jizhen Military Road: Branch	Yiyuankou, Yipianshi, Luowenyu	Length of the border; Number of attached walls
Jizhen Defensive Unit: Pass	Shanhaiguan; Yiyuankouguan; Jianganlingguan	Altitude; Slope; longitude; latitude
Jizhen Defensive Unit: Zhaibao	Wutonggubao; LiuheWeibao; Xinkailingbao	Altitude; Slope; longitude; latitude
Jizhen Defensive Unit: Camp	Taitou camp; Santun camp; Fumazhai camp	Altitude; Slope; longitude; latitude
Jizhen Logistics Unit: Granary	Longqingcang; Yongyingcang; Guangchucang	Station
Jizhen Logistics Unit: Relay	Yuguan Station; Luanyang Station; Yuyang Station	Station
Folklore settlement: Township	Xiaotangshan Town; Yanchi Town; Dazhuangke Township	District/County
Folklore settlement: Village	Lingjiao Village; Huaguoshan Village; Xiangtang Village	Township; Introduction
Folklore settlement: Tourist Village	Baiyu Village; Tianlongtan Village; Lianhuachi Village	Township; Introduction
War event	Battle of Jingnan; Tumubao Mutiny	Time; Result; Introduction; Place of incident
District County	Yanqing; Changping; Qianxi County; Shanhaiguan District	Location; Introduction
Ethnic organization	Tatar; Wala; Khitan	Alias; Introduction
Weapon	Farangi;Three-Eyed Gun; Iron General Cannon	Aliases; Brief history

TABLE II

DETAILS OF ALL RELATIONS IN THE ONTOLOGY LAYER

Relation	*Instance*
Initial time	Building, Official, Logistics Unit ⇒ Time
Dynasty of the book	Historical works ⇒ Time
Incident time	War event⇒Time
Time of general soldier	Chief Military Officer⇒Time
Place of birth	Person⇒Sanitation, District County
Place of incident	War event⇒Jizhen Defensive Unit
Affiliated area county	Scenic spot, Sanitation⇒District County; Military Road, Township⇒District County
Garrison location	Official⇒Military Road, Defensive Unit
Exclusive support	The Guerrilla⇒ Military Road
Immediate boss	Official⇒Official
Reigning emperor	Reign Title⇒Emperor
Officer stationed	Military Road⇒Official
Official position	Person⇒Official
West side	Nine Sides⇒Nine Sides; Military Road⇒ Military Road
East side	Nine Sides⇒Nine Sides; Military Road⇒Military Road
Father	Person⇒Person
Son	Person⇒Person
Uncle	Person⇒Person
Grandfather	Person⇒Person
Grandson	Person⇒Person
Brother	Person⇒Person
Town belongs	Military Path⇒Nine Sides
Military road belongs	General Road⇒Military Path; Branch⇒General Road
Township belongs	Village⇒Township
Branch belongs	Defensive Unit, Logistics Unit⇒Branch
Author	Books⇒Person
Plants	Scenic spot⇒Vegetation
Main building	Scenic spot⇒Building
Builder	Defensive Unit⇒Person
Main participants	War event⇒Person,Ethnic organization

updated and improved, and it is bound to face the growth of relation types in the new corpus.

For these reasons, we choose to adopt the Semi-supervised Relational Siamese Network (RSN) [3] proposed by Ruidong Wu and Yuan Yao to learn from both supervised data of pre-defined relations and unsupervised data with novel relations. It is an Open Relation Extraction (OpenRE) method based on clustering. Specifically, Semi-supervised RSN learns a relation similarity metric from labeled data of pre-defined relations and then transfers the metric to measure the similarity of unlabeled sentences for open relation clustering. We denote a small number of labeled training data sets as d_l and unlabeled training data sets as d_u. The architecture of Semi-supervised RSN is shown in Fig 2, including the relation similarity calculation module, loss module, and relation clustering module.

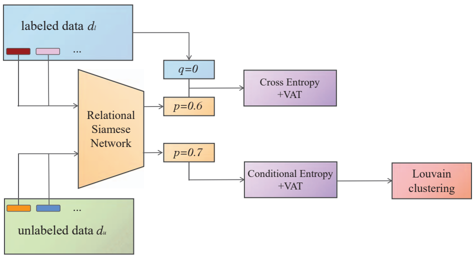

Fig. 2. The technical architecture of Semi-supervised RSN.

RSN predicts whether two sentences mention the same relationship, namely the relation similarity. The architecture of RSN is shown in Fig 3. The input layer represents the sentence $\boldsymbol{x}$ containing $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$ entities as pre-trained word embeddings and randomly initialized position embeddings. The CNN layer comprises multiple shared convolutional layers and a maximum pooling layer to compress the input sequence. Finally, the FC layer with sigmoid activation maps the feature to the relation vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. The similarity computation layer converts the absolute distance of the two relation vectors outputted by the CNN layer into relation similarity $p\epsilon\,[0,1]$. This layer is a one-dimensional output FC layer with sigmoid activation. The calculation expression of p is as follows:

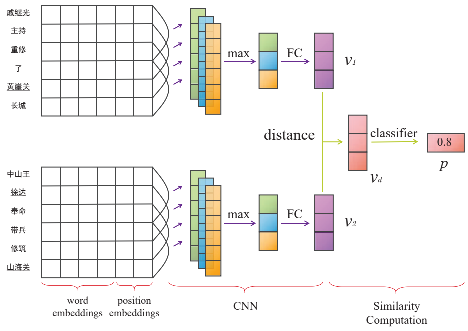

Fig. 3. The technical architecture of RSN.

$$p = \sigma \left(\omega \left| CNN \left(\boldsymbol{x}_1 \right) - CNN \left(\boldsymbol{x}_2 \right) \right| + b \right),$$

where σ is the sigmoid function, and w and b are weights and deviations. The two CNN modules are identical and share all the parameters.

For the labeled data set d_l, the training of the RSN is based on the cross-entropy loss by the binary label q and relationship similarity p as follows:

$$\pounds_l = \mathbb{E}_{d_l} \left[q \, ln \left(p_\theta \left(d_l \right) \right) + \left(1 - q \right) ln \left(1 - p_\theta \left(d_l \right) \right) \right],$$

where θ is all the parameters in RSN.

The conditional entropy loss is an effective method for optimizing the classification edge, which penalizes the close-boundary distribution of data points by maximizing the distances from the relational instances to the decision boundary:

$$\pounds_u = \mathbb{E}_{d_u} \left[p_\theta \, ln \left(p_\theta \left(d_u \right) \right) + \left(1 - p_\theta \left(d_u \right) \right) ln \left(1 - p_\theta \left(d_u \right) \right) \right].$$

Virtual adversarial loss can search the neighborhood of data points and penalize the most dramatic changes in distance prediction, thereby effectively alleviating over-fitting and improving the generalization ability of the model:

$$\pounds_{vl} = \mathbb{E}_{d_l} \left[D_{KL} \left(p_\theta \left(d_l \right) \parallel p_\theta \left(d_l, t_1, t_2 \right) \right) \right],$$

$$\pounds_{vu} = \mathbb{E}_{d_u} \left[D_{KL} \left(p_\theta \left(d_u \right) \parallel p_\theta \left(d_u, t_1, t_2 \right) \right) \right],$$

where D_{KL} indicates the Kullback-Leibler divergence, and $p_\theta \left(d_l, t_1, t_2 \right)$ indicates a new distance estimation with disturbances t_1 and t_2 on the two input instances. Empirically, the approximate perturbation is the same as the original paper [16]. In summary, Semi-supervised RSN is trained by the following joint loss function, learning from labeled and unlabeled data:

$$\pounds_{all} = \pounds_l + \lambda_v \pounds_{vl} + \lambda_u \left(\pounds_u + \lambda_v \pounds_{vu} \right),$$

where λ_v and λ_u are hyper-parameters.

The semi-supervised RSN clusters the target relation instances of the new relation type by Louvain [17] algorithm in the relation clustering module. Louvain is a graph-based clustering algorithm traditionally used to detect communities. This algorithm automatically finds the appropriate cluster size by optimizing the community modularity without predetermining the number of potential clusters. Finally, we reviewed and revised the new relation cluster under the guidance of experts.

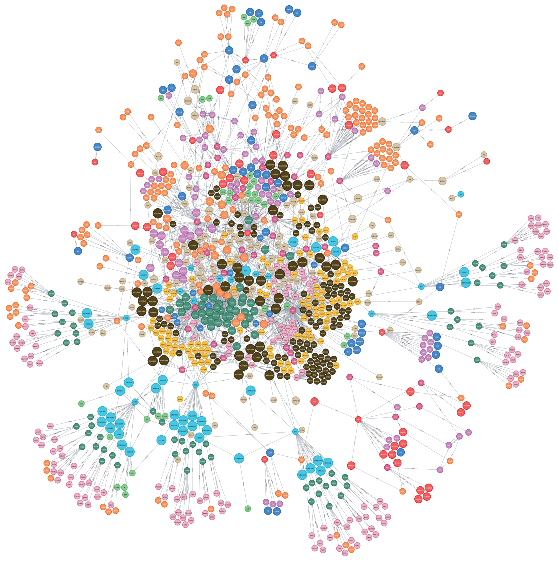

Fig. 4. Visualization results of GWKG utilizing Neo4j.

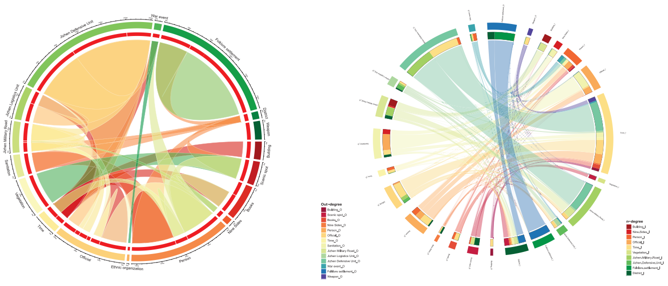

Fig. 5. Visualization results of GWKG utilizing Circos.

D. Attribute Filling

Attribute filling can enhance the semantic interpretation of entities, enrich the knowledge density and practical value of GWKG's data layer, and promote subsequent in-depth data analysis and knowledge reasoning. We first link the entities to Baidu Baike and handle them as seeds, then iteratively crawl online web pages. Finally, we only need to analyze the information block format and write the triggers to complete it.

The entity linking associates the entities extracted by NER with the entities in the knowledge base. In big Internet data, encyclopedia knowledge bases contain rich entity information and are often organized in a particular structure. It is very convenient to extract entity attributes from such data sources. Linking entities in the KG to the entities extracted by NER has become a trend.

Web crawlers are vital for the search engine crawling system. The basic workflow of web crawlers is first to take the URL of the encyclopedia webpage of the entity link as the collected seed, second to adopt a depth-first traversal strategy, starting from the seed start page, tracking downlink by link and downloading to the local machine. We exploit a trigger based on linguistic rules to automatically extract the entity attribute value of a given attribute name. Encyclopedia pages usually consist of a series of information blocks. Each information block has its unique location and is wrapped by a specific tag. We only need to write a rule trigger for the information block format on the page to complete the attribute extraction.

IV. Results

We chose CNKI and the literature library of the cultural industry as the original data and then relate Baidu Baike for data enhancement to build GWKG. It contains 34 concepts, 33 types of relationships, more than 6,000 entities and attributes, and 720,000 Chinese corpora. We explore various presentation forms to visualize GWKG, include Neo4j, Circos, Gephi, and D3. Due to space limitations, this section only shows the visualization results of Neo4j and Circos. Other results and the GWKG data set, including CSV, JSON, and RDF files, can be found on https://github.com/Feng0204/GWKG.

Neo4j is one of the existing mainstream native graph databases, often adopt to store KGs. Fig 4 shows the visualization of GWKG through Neo4j. GWKG contains the semantic relations between various types of Great Wall entities. The concept of the node is filled with different colors, for example, yellow represents the Pass, and pink represents the Tourist Village. It is not difficult to see the high density of GWKG from Fig 4, reflecting the advantages of the semi-automatic construction of KGs. For cultural heritage workers, they can reason from GWKG through the Neo4j interface to master the knowledge at multiple levels of concepts, entities, relations, and attributes, which can effectively improve searching and organizing information.

Circos [4] visualizes types of positional relationships between genomic intervals. These data are usually generated by sequence alignment, hybridization arrays, genome mapping, and genotyping studies. Inspired by this, we utilize Circos to visualize the semantic relations in GWKG. As far as we know, this is the first time that Circos has been applied to visualize KGs. The left of Fig 5 shows the correspondence between different concepts. The outermost circle represents 17 coarse-grained concepts, the second red circle represents 34 fine-grained concepts, and the arc length represents the number of entities contained under the concept. The 33 types of relationships are displayed in the middle ribbon, making it easy to identify the corresponding relationships between different concepts. The right of Fig 5 is an abundance map of GWKG, presenting the out-degree and in-degree of each relationship. The width of the middle color band represents the number of corresponding relationships.

V. Conclusion

In this paper, we introduce the construction of a semi-automated knowledge graph GWKG. It covers a high-quality fine-grained ontology layer that can facilitate knowledge inference and analysis in the heritage domain. In the future, we hope to integrate and link more Great Wall cultural data to GWKG. We will continue to maintain and update GWKG to provide more technical support for the protection of the Great Wall culture heritage.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT*, 2019.

[2] Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents,” *ArXiv*, vol. abs/1405.4053, 2014.

[3] R. Wu, Y. Yao, X. Han, R. Xie, Z. Liu, F. Lin, L. Lin, and M. Sun, “Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 2019.

[4] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. M. Jones, and M. Marra, “Circos: an information aesthetic for comparative genomics.” *Genome research*, vol. 19 9, pp. 1639–45, 2009.

[5] S. Zheng, J. Rao, Y. Song, J. Zhang, X. Xiao, E. Fang, Y. Yang, and Z. Niu, “Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining.” *Briefings in bioinformatics*, 2020.

[6] D. Wishart, Y. D. Feunang, A. Guo, E. J. Lo, A. Marcu, J. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, “Drugbank 5.0: a major update to the drugbank database for 2018,” *Nucleic Acids Research*, vol. 46, pp. D1074 – D1082, 2018.

[7] P. Ernst, C. Meng, A. Siu, and G. Weikum, “Knowlife: A knowledge graph for health and life sciences,” in *IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014*, 2014.

[8] R. Page, “Ozymandias: A biodiversity knowledge graph,” *bioRxiv*, 2018.

[9] G. Delacruz, G. K. W. K. Chung, and W. L. Bewley, “Characterizing trainees in the cognitive phase using the human performance knowledge mapping tool (hpkmt) and microgenetic analysis,” 2006.

[10] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and B. Yang, “Knowedu: A system to construct knowledge graph for education,” *IEEE Access*, vol. 6, pp. 31 553–31 563, 2018.

[11] M. Cheatham, A. A. Krisnadhi, R. Amini, P. Hitzler, K. Janowicz, A. Shepherd, T. Narock, M. B. Jones, and P. Ji, “The geolink knowledge graph,” *Big Earth Data*, vol. 2, pp. 131 – 143, 2018.

[12] V. Maltese and F. Farazi, “A semantic schema for geonames,” 2013.

[13] P. Haase, D. M. Herzig, A. Kozlov, A. Nikolov, and J. Trame, “metaphactory: A platform for knowledge graph management,” *Semantic Web*, 2019.

[14] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, and C. Veninata, “Arco: The italian cultural heritage knowledge graph,” in *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, ser. Lecture Notes in Computer Science, 2019.

[15] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *ArXiv*, vol. abs/1706.03762, 2017.

[16] T. Miyato, A. M. Dai, and I. Goodfellow, “Adversarial training methods for semi-supervised text classification,” *arXiv: Machine Learning*, 2017.

[17] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. 10008, 2008.