From Text to Historical Ecological Knowledge: The Construction and Application of the *Shan Jing* Knowledge Base

Ke Liang[1], Chu-Ren Huang[1], Xin-Lan Jiang[2]
[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
[2] Department of Computer Teaching and Research, University of Chinese Academy of Social Sciences
[1]11 Yuk Choi Rd, Hung Hom, Hong Kong
[2]11 Changyu Rd, Liangxiang Higher Education Park, Fang Shan, Beijing
leo-ke.liang@connect.polyu.hk, churen.huang@polyu.edu.hk, jxlzgl@sina.com

Abstract

Traditional Ecological Knowledge (TEK) has been recognized as a shared cultural heritage and a crucial instrument to tackle today's environmental challenges. In this paper, we deal with historical ecological knowledge, a special type of TEK that is based on ancient language texts. In particular, we aim to build a language resource based on *Shanhai Jing* (*The Classic of Mountains and Seas*). Written 2000 years ago, *Shanhai Jing* is a record of flora and fauna in ancient China, anchored by mountains (shan) and seas (hai). This study focuses on the entities in the *Shan Jing* part and builds a knowledge base for them. We adopt a pattern-driven and bottom-up strategy to accommodate two features of the source: highly stylized narrative and juxtaposition of knowledge from multiple domains. The PRF values of both entity and relationship extraction are above 96%. Quality assurance measures like entity disambiguation and resolution were done by domain experts. Neo4j graph database is used to visualize the result. We think the knowledge base, containing 1432 systematically classified entities and 3294 relationships, can provide the foundation for the construction of a historical ecological knowledge base of China. Additionally, the ruled-based text-matching method can be helpful in ancient language processing.

Keywords: *Shan Jing*, Traditional Ecological Knowledge, knowledge base, ancient language processing

1. Introduction

Traditional Ecological Knowledge (TEK) refers to the inherited knowledge of the relationships between living beings in a specific ecosystem. TEK encompasses vital information about the Earth's evolutionary history and the development of human civilization and can be derived from original holders and oral histories (McGregor, 2008; Mustonen, 2019). Integrating TEK with scientific knowledge can assist us in addressing environmental challenges (Gagnon & Berteaux, 2009). Historical ecological knowledge is a special type of TEK. As an initial work in creating China's historical ecological knowledge base, this paper focuses on the historical ecological knowledge in *Shan Jing* and tries to build a knowledge base for it.

Shan Jing 山经 is a part of the ancient Chinese book *Shanhai Jing* 山海经 (*the Classic of Mountains and Seas*) which was written in the pre-Qin period. Together with *Hai Jing* 海经 (*the Classic of Seas*) and *Huang Jing* 荒经 (*the Classic of the Greater World*), it constitutes *Shanhai Jing* seen today. Containing nearly 20,000 words, *Shan Jing* preserves a significant amount of flora, fauna as well as geographical information two millennia ago. Its abundance of domain-specific entities, along with its unique narrative language, makes the knowledge extraction quite different from the prevalent methodologies in the general knowledge domain of modern Chinese. Based on a detailed analysis of the *Shan Jing* text, we employ a rule-based method to extract entities in *Shan Jing* and classify them into a 15-category vocabulary list by referring to the annotations about *Shan Jing* which were made by scholars from previous generations. The vocabulary list serves as the foundation for knowledge base construction. The entire construction process involves digital text preprocessing, entity and relationship extraction, entity disambiguation and resolution, and knowledge base visualization with Neo4j.

The research method and results outlined in this paper hold several practical applications. Firstly, the knowledge base containing systematically classified entities and relationships can provide a foundation for the construction of a historical ecological knowledge base of China, and for the construction of *Shan Jing*'s monograph intelligent book, as well as assisting scholars in their studies about *Shan Jing*. Furthermore, the rule-based text-matching approach can be employed for the collation of ancient books and other related tasks in ancient language processing.

The reasons why this first study focuses only on *Shan Jing* are the following. Historically, *Shan Jing*, *Hai Jing*, and *Huang Jing* were considered as one book (Shanhai Jing), to be studied and understood together since at least the Han Dynasty (Chen, 2012). However, since the study by Bi Yuan 毕沅 of the Qing Dynasty, there have been ongoing controversies over the relationship between the three parts (Chen 2012). For instance, Liu Zongdi (2006) suggests that *Shan Jing* is a written record of on-site surveys, while *Hai Jing* and *Huang Jing* describe images in words. These controversies imply that *Shanhai Jing* should not be taken for granted as a unified whole. Another reason is due to operational considerations. *Shan Jing* shows a logical and highly stylized narrative, while the narratives of *Hai Jing* and *Huang Jing* are relatively ambiguous. Since *Shan Jing* has the best textual coherence and is often considered to be the most reliable empirical account, it is a natural starting point for the construction of a knowledge base. It is expected that the structure of the knowledge base and methodologies developed here can be applied later to *Hai Jing* and *Huang Jing*.

LREC-COLING 2024, pages 7521–7530

2. Related Work

Guo Pu 郭璞 (276-324 CE) authored the first annotation of *Shanhai Jing* and heralded a long history of studies on the geology and living beings described in this book. The almost mythical world in *Shanhai Jing* has great appeal not only to ancient scholars such as Zhu Xi 朱熹 and Hao Yixing 郝懿行 but also to scholars today. For example, Woo's (2002) study on Taoism and the ecology of ancient China contained a lot of discussion involved with *Shanhai Jing*. Guo (2004) attempted to find realistic prototypes for the living beings described in *Shanhai Jing* from a biological perspective. Strassberg (2018) merged translations from the original work with information on archaeological discoveries to present the fascinating pageant of strange creatures to readers. Furthermore, Fu (2021) considered *Shanhai Jing* as the origin of the modern ecological narrative, which can help us recall long-lost ecological memories. This paper focuses on analyzing the ecological knowledge in this book from a lexical perspective. In this regard, we would like to give a brief introduction to related work from two aspects: ancient Chinese historical ecological knowledge engineering, and research about *Shanhai Jing*'s lexical items.

The construction of the ancient Chinese historical ecological knowledge base can be traced back to the early 2000s. when Huang et al. (2004) constructed a domain ontology for Tang poems. They adopted the Shakespearean garden approach and manually tagged lexicons for the categories of animal, plant, and artifact, and mapped these historical floral and faunal lexicons to the English-Chinese bilingual wordnet system, Sinica BOW. Similarly, Hsieh et al. (2006) employed a lexicon-driven and ontology-merging methodology to incorporate plant names from Guang-Qun-Fang Pu 广群芳谱 into Sinica BOW. They conducted a meticulous analysis of the botanical pedigrees in this book. In recent years, machine learning and deep learning techniques have also been applied to this area. Wu et al. (2023) constructed a Chinese named entity recognition model for plants using the CRF algorithm and various deep-learning models. Based on their model, a knowledge graph for plant knowledge in pre-Qin classics was built. Although Wu's work may not fully meet the meticulous requirements of humanities scholars for textual research, the technical approaches they employed provide us with a certain reference for the subsequent construction of China's historical ecological knowledge base.

Han Yiying 韩一鹰 (1933; 1936) was the first to research the lexicons of *Shanhai Jing*. In these two papers, he analyzed the names of animals, plants, and deities in *Shanhai Jing* based on manually collected information and qualitative studies. Yuan Ke 袁珂 should be credited as the first scholar to systematically analyze the lexical items of *Shanhai Jing*. In his book *Shan-hai-jing-jiao-zhu* 山海经校注 (Yuan 1980), he cataloged over 4000 instances of proper names in *Shanhai Jing* and arranged them according to stroke order of the first Chinese character. In the 1990s, Korean scholar Kyung Ho Suh (1993) presented the lexical items of *Shan Jing*, *Hai Jing*, and *Huang Jing* in table format in the appendix of his doctoral dissertation. These tables resemble two-dimensional tables in relational databases, which depict the connections between different items. Although both Yuan Ke and Kyung Ho Suh conducted relatively comprehensive research on the lexicon in *Shanhai Jing*, they focused on inventory rather than classification. This gap was addressed by Jia (2004) in his doctoral thesis. In Jia's thesis, titled *Research on Proper Names of Shan-hai-jing*, he focused on the proper names as the main research objects, dividing them into 18 categories, such as deity's name, prefecture's name, and mountain's name, among others. In the initial stages of our paper, Jia's research served as an important reference for our classification of the lexical system of *Shan Jing*. However, since our research focuses on building a coherent knowledge system based on *Shan Jing*, our taxonomy may deviate slightly from Jia's.

The rich historical ecological knowledge in *Shanhai Jing* provides us with a window to understanding ancient culture. Moreover, by revisiting ancient history, we can better comprehend our relationship with nature and effectively address present-day environmental challenges. However, traditional humanities research has primarily focused on the lexical analysis of the knowledge within *Shanhai Jing*, with limited consideration given to its expansion and application. In the current wave of digital humanities, we utilize computational methods to organize and categorize the historical ecological knowledge in *Shan Jing* (which will be expanded to the entire *Shanhai Jing* in the future) and visualize it via the Neo4j graph database. We hope that our research can fill the gap in historical ecological knowledge engineering in the field of *Shanhai Jing* studies, and we also hope that it can eventually form a Chinese historical ecological knowledge base together with other ancient Chinese historical ecological knowledge engineering research.

3. Pattern-Driven and Bottom-Top Methodology

Figure 1 depicts the overall workflow of this paper, which can be divided into two main parts: schema design and knowledge base construction. It is worth noting that there are two extraction tasks in the workflow. One is in the Schema Designing part and the other one is in the Knowledge Base Construction part. In the schema designing phase, the instance extraction aims to collect instances from the text which will be manually categorized later. In the knowledge base construction phase, entity and relationship extraction is performed to transform unstructured text into structured knowledge, facilitating the construction of a knowledge graph. The categorization results from the first extraction will be utilized in the second extraction. Both extractions follow the same syntactic patterns, but the algorithmic logic of the latter extraction is more complex than the former. This section primarily focuses on the design of the schema.

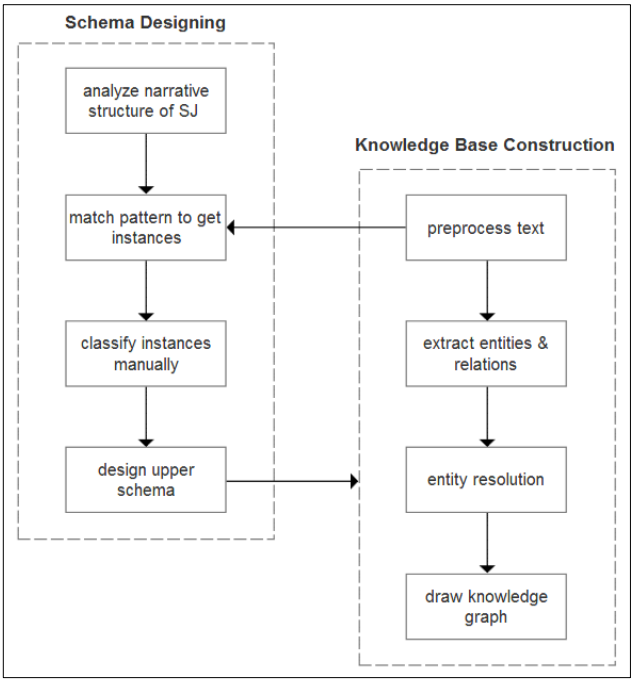

Figure 1: The workflow of design and construction.

3.1 The Narrative Structure of *Shan Jing*

Shan Jing is comprised of five parts, which are the *Classic of Southern Mountains* (consisting of 3 chapters), *Classic of Western Mountains* (4 chapters), *Classic of Northern Mountains* (3 chapters), *Classic of Eastern Mountains* (4 chapters), and *Classic of Central Mountains* (12 chapters). Altogether, these five parts consist of a total of 26 chapters. In each chapter of every part, *Shan Jing* follows a consistent narrative pattern. The three passages[1] presented in Figure 2 are the first three paragraphs of the *First Classic of Southern Mountains*. In the introduction of landmarks and their associated products, certain formulaic expressions are used repeatedly. In this figure, colored highlights are used to mark these formulaic patterns, with different colors assigned to different patterns.

南山经之首曰鹊山。其首曰招摇之山，临于西海之上，多桂，多金、玉。有草焉，其状如韭而青华，其名曰祝余，食之不饥。有木焉，其状如谷而黑理，其花四照，其名曰迷榖，佩之不迷。有兽焉，其状如禺而白耳，伏行人走，其名曰狪狪，食之善走。丽麃之水出焉，而西流注于海，其中多育沛，佩之无瘕疾。

又东三百里，曰堂庭之山，多棪木，多白猿，多水玉，多黄金。

又东三百八十里，曰猨翼之山，其中多怪兽，水多怪鱼，多白玉，多腹虫，多怪蛇，多怪木，不可以上。

Figure 2: The passages in *Shan Jing*.

From these three paragraphs, we can get a general understanding of the narrative style of *Shan Jing*. By further analyzing the first chapter of each of the five sections (the Classic of Southern/ Western/ Northern/ Eastern/ Central Mountains), we can summarize the following schemas:

又{方向}{距离}里，曰{山名}之山，临于/望于/鎜于{地名}，多/无{物产名}。有{物产类别}焉，（其状如……，其音如……），其名曰{物产名}。{水名}出焉，而{方向}流注于{水名}，其中多/无{物产名}……。[2]

{distance} leagues further {direction} is a mountain called Mount {mountain name}, is close to {landmark name}.It has lots of/has no {resource name}. There is a kind of {resource category} on this mountain (which looks like……, which sounds like……). Its name is {resource name}. The {river name} originates from here and flows {direction} to empty into {river name}. It has lots of/has no {resource name}…….

In the above schemas, each pair of curly brackets contains either a named entity or a property assigned to a particular named entity, as annotated above. Applying these schemas, the narrative structure of a particular chapter consists of a series of paragraphs, each beginning with an introduction to a mountain. Within each paragraph, it contains several schemas describing the entities associated with the mountain and their properties. Rivers, in particular, are

[1] The meaning of these three paragraphs is :

The first mountain range in the Classic of the Southern Mountains is called Mount Que4. Its first peak is called Mount Zhao1-Yao2. It is close to the West Sea. It has lots of cinnamon trees, gold and jade. There is a kind of grass on this mountain which looks like an onion, but it has green flowers. Its name is Zhu4-Yu3. If you eat it, you won't starve. There is a kind of tree on the mountain which looks like the paper mulberry, but it has black markings. Its blossoms light up everything around it. Its name is Mi2-Gu3. If you wear it in your belt, you won't get lost. There is a kind of animal on the mountain which looks like a long-tailed ape, but it has white ears. It crouches as it moves along and it runs like a human. Its name is Sheng1-Sheng1. If you eat it. You'll be a good runner. The River Li4-Ji3 originates from here and flows westwards to empty into the sea. It has lot of Yu4-Pei4, if you wear some in your belt, you won't suffer from worms.

Three hundred leagues further east is a mountain called Mount Tang2-Ting2. It has lots of fruiting shrubs, white gibbons, rock crystal and yellow gold.

Three hundred and eighty leagues further east is a mountain called Mount Yuan1-Yi4. It has lots of beasts, and many curious fish in its waterways. It has lots of white jade, giant snout-snakes, strange serpents and curious trees. You can't climb up this mountain.

[2] Except for the pattern of "又{方向}{距离}里，曰{山名}之山" ({distance} leagues further {direction} is a mountain called Mount {mountain name}), which appears only once in a paragraph, the other patterns may be repeated multiple times within the same paragraph. And the pattern of "临于/望于/鎜于{地名}" (is close to {landmark name}) is just a simplified summary, and such pattern in *Shan Jing* text may also involve descriptions of directions.

described in terms of the mountain that originates from, its destination, and the direction in which the river flows, as well as the resources the river has. Occasionally, it also mentions other landmarks after the mountain, such as the West Sea in the "Mount Que4" paragraph. Representing such information in the text using triplets, we can identify the following relationship types:

- The relationship between mountains: (Mount A, direction, Mount B).
- The relationship between mountain and river: (River A, originates from, Mount B).
- The relationship between rivers: (River A, direction+flows into, River B).
- The relationship between the mountain and its resources: (Mount A, has (lots of)/ has no, [animals, plants, minerals et al.]).
- The relationship between the river and its resources: (River A, has (lots of)/ has no, [animals, plants, minerals et al.]).
- The relationship between mountain and other landmarks: (Mount A, is close to, landmark B)

For the fourth and fifth relationship, the animals and plants within the product resource can be further classified into more specific categories. For instance, the plant species mentioned in *Shan Jing* can be categorized into two subcategories: trees and grasses.

3.2 Pattern-based Instance Extraction

Based on the analysis of *Shan Jing*'s syntactic structures, we have designed corresponding regular expressions to extract instances from them. Then these instances will be manually classified into different vocabularies.

As the narrative patterns in section 3.1 show, in *Shan Jing,* each entity is anchored by a specific syntactic pattern, which is usually a verb or verb phrase but can also be a prepositional phrase. In this paper, we mainly use the six patterns in Table 1 to conduct the corresponding entity extraction (relationship extraction as well).

No.	Syntactic pattern	Instance
1	又{方向}{距离}里，曰{山名}之山 {distance} leagues further {direction} is a mountain called Mount {mountain name}	mountain
2	{水名}出焉 The {river name} originates from here	river
3	而{方向}流注于{水名} flows {direction} to empty into {river name}	river
4	多/无{物产名} has lots of/has no {resource name}	resource
5	有{物产类别}焉，……，其名曰{物产名} There is a kind of {resource category} on this mountain, ……. Its name is {resource name}	resource
6	临于/望于/鎮于{地名} close to {landmark name}	another landmark

Table 1: The syntactic patterns for extraction.

We can see that for instances of mountains and other landmarks, we only need to use one syntactic pattern to do the extraction. For the instances of river and resource, we need to use two rules to extract them respectively. This is because when introducing rivers, different descriptions are used for the origin and endpoint of the river; when introducing resources, *Shan Jing* provides both detailed and concise descriptions.

By utilizing regular expressions alone, we can extract a significant portion of instances from *Shan Jing*. These instances can then be checked with *Shan Jing's* text to do further categorization.

3.3 Vocabulary Building and Upper-level Schema Designing

The instances extracted in section 3.2, especially those belonging to product resources, can be further divided into different categories. We classify these resources such as grass, birds, fish, and so on, ultimately dividing the instances from the *Shan Jing* into 15 different categories. Then we build vocabulary for each category and assign alphabetical tags to it, as shown in Table 2. These category tags will be used in the construction of the knowledge graph.

Tag-category				
A-mountain	B-river	C-grass	D-tree	E-bird
F-beast	G-snake	H-mineral	I-fish	J-shell
K-turtle	L-other aquatic life	M-deity	N-other landmark	O-other

Table 2: Tag-category pair.

Based on the analysis of *Shan Jing*'s narrative structure and its category system, we design the top schema for future knowledge base construction, which is shown in Figure 3. In this figure, the rectangles represent entities and the arrows represent relationships. To make the final knowledge base also show the chapter information, we also added the Classic name 经名 and the chapter information (the item in the dotted boxes) in addition to the entities mentioned above. The items in the solid boxes and the associated relationships correspond completely to the triplets mentioned above.

4. The Construction of Knowledge Base

As Figure 1 shows, the construction of a knowledge base involves four main parts: text preprocessing, entity and relationship extraction, entity disambiguation and resolution, and knowledge base visualization.

4.1 Dataset and Preprocessing

The digital text used in this study was obtained from the Chinese Text Project[3] digital library. It comes from the *First Edition of the Four Series* 四部丛刊初编, which was written during the Chenghua period of the Ming Dynasty. Since this paper employs a rule-based method for knowledge extraction, there is no need to do word segmentation. The text preprocessing mainly involves correcting errors in the digital text. We proofread the digital text with its original photocopy and Yuan Ke's *Shan-hai-jing Full Translation* 山海经全译 (2016), making necessary corrections to the identified errors in the text. The collated version of *Shan Jing* contains 26,681 Chinese characters (including punctuation), 1,445 sentences, and 481 paragraphs.

It is noteworthy that ancient Chinese literature lacks punctuation, and the punctuation seen in *Shan Jing* today was added by scholars in later periods. As punctuation plays a crucial role in knowledge extraction, we primarily relied on Yuan Ke's *Shan-hai-jing Full Translation* to rectify the inappropriate punctuation in the text. For instance, following Yuan Ke's book, we revised "多金玉" (has lots of gold and jade) to "多金、玉".

4.2 Entity and Relation Extraction

The entity and relation extraction process is similar to the previously mentioned instance extraction process: the same syntactic patterns (see Table 1) are used, but this time we also have to pay attention to the relations that were ignored in the instance extraction.

The verb/prepositional phrases in Table 1, which we used as patterns to extract instances also convey the relationships between different entities. For example, "多" (has lots of) and "流注于" (flows into) not only aid in locating entities but also indicate the semantic relationship between entities. By focusing mainly on these phrases, we can extract entities and relationships simultaneously.

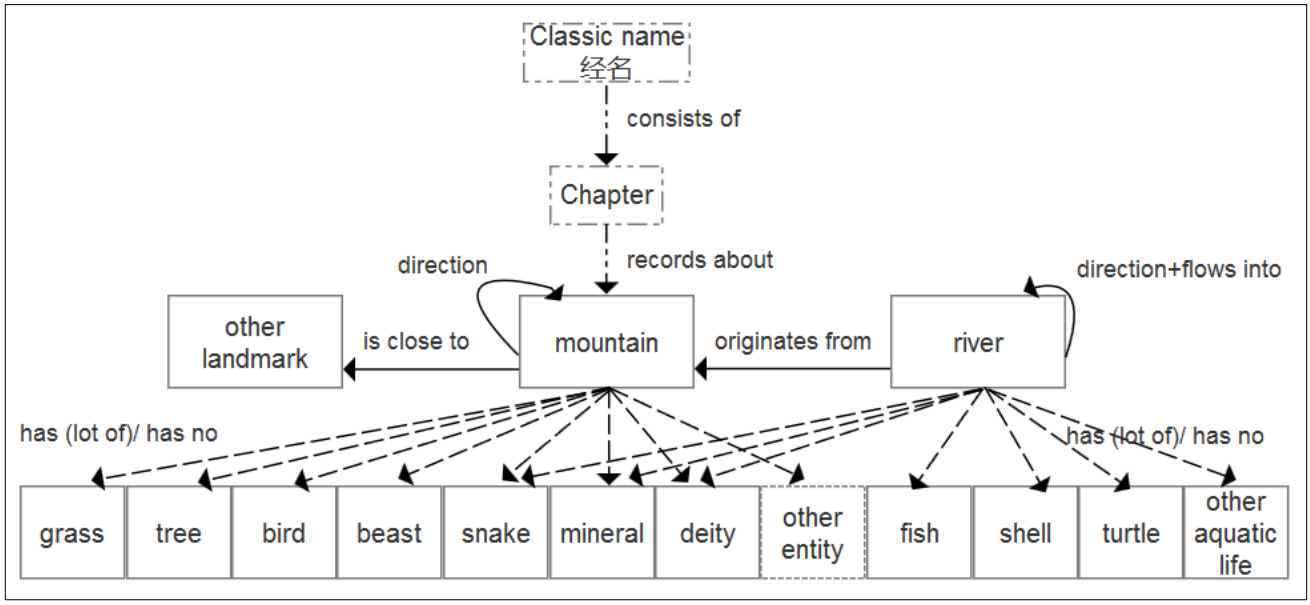

Figure 3: The schema of knowledge base.

The extraction process can be seen as a transformation of unstructured text data into fully structured tabular data. In addition to the regular expressions, an algorithm is required to accurately store the entities and their relationships. The algorithm presented in Table 3 employs a loop structure that takes paragraphs as larger units and sentences as smaller units to extract entities and relationships. The extracted results from the same passage are stored in the same row, with the entities and relationships categorized and stored in corresponding columns. To avoid confusion between entities and their relationships, the category tags in Table 2 are added as prefixes to the entities during storage. Additionally, for entities associated with product resources, a numeric tag is added to distinguish different "has" relationships (多 has lots of – 0, 无 has no – 1, 有 there is a kind of - 2).

Algorithm: knowledge extraction and storage

Input: *Shan Jing* text, vocabulary list

Output: Similar-relational database

define different column indexes to different entity/ relationship categories

for paragraph i in *Shan Jing's* text:

  for sentence j in paragraph i:

    extract entities and directions in the sentence

    check the category of the entity

    add category-tag as a prefix to the entity

    if the entity is a product:

      check the verb in the corresponding

[3] https://ctext.org/shan-hai-jing/zh

9 add has-tag (has/ has lots of/ has no) as a prefix to the entity

10 **end**

11 **end**

12 write the entities or directions of the same paragraph in the same row, and write entities of different types to different column indexes defined before

13 **end**

Table 3: The algorithm for extraction and storage

We finally extract 1401 different entities from the *Shan Jing* text, and their distribution in each category is shown in Table 4. It can be observed that the number of entities such as mountains or rivers is the largest, followed by plants, birds, and beasts.

Category	Number	Category	Number
Mount	461	Fish	49
River	301	Shell	10
Grass	74	Turtle	8
Tree	105	Other aquatic life	9
Bird	76	Deity	36
Beast	128	Plain	25
Snake	16	Others	9
Mineral	94	**SUM**	1401

Table 4: The distribution of entities.

We extract a total of 2802 relationships, and their distribution among the six relationship types is shown in Table 5. We can find that the number of relationships between mountains and their resources accounts for a significant portion of the total. This aligns with the fact that the *Shan Jing* text dedicates considerable content to describing the resources associated with mountains.

Relationship type	Number
(Mount A, direction, Mount B)	422
(River A, originates from, Mount B)	258
(River A, direction+flows into, River B)	246
(Mount A, has (lots of)/has no, resource)	1648
(River A, has (lots of)/has no, resource)	184
(Mount A, is close to, landmark B)	44
SUM	2802

Table 5: The distribution of relationships.

As mentioned earlier, we designed extraction patterns based on the analysis of the first chapter of each section in *Shan Jing*. After completing the extraction, we manually review all of them by referring to the original text. Therefore, we can say that our test set consists of the entire text of *Shan Jing*. Due to the uneven distribution of entities and relationships in different categories, we chose to measure the performance of the model using Micro-PRF. We first calculate the sum of true positives (TP), false positives (FP), and false negatives (FN) for all categories in the entity confusion matrix and relationship confusion matrix, then get the averages of these sums. Finally, we use these averages to calculate the overall precision, recall, and F1 score. The calculation formulas are as follows:

$$micro - P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}$$

$$micro - R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}}$$

$$micro - F1 = \frac{2 \times micro - P \times micro - R}{micro - P + micro - R}$$

Table 6 depicts our final performance data. Both the precision and recall rates are above 96%, indicating that the pattern-driven method is well-suited for initial digital humanities studies on *Shan Jing*.

Symbol	Precision	Recall	F1-score
Entity Extraction	96.421%	96.353%	96.387%
Relation Extraction	96.516%	96.895%	96.705%

Table 6: The measurement of entity and relationship extraction result.

4.3 Entity Disambiguation and Resolution

Entity disambiguation aims to resolve the ambiguity of entities that have the same or similar names but represent different concepts or entities in the real world. Entity resolution, on the other hand, focuses on identifying mentions that are not the same but refer to the same concept or real-world entity. In *Shan Jing*, some entities share the same name but denote different things, while other entities with different names refer to the same object. Therefore, the tasks of entity disambiguation and entity resolution become necessary. Given that in *Shan Jing's* text determining the correct referent of these entities requires expert knowledge, entity disambiguation, and entity resolution are performed manually.

Let's discuss entity disambiguation first. In *Shan Jing*, entities with the same name but referring to different things can be divided into two types. The first type (type A) includes entities that have the same name and belong to the same category, mainly mountains. For example, "丙山" (Mount Bing3) appears in both the *Tenth Classic of Central Mountains* and the *Twelfth Classic of Central Mountains* but refers to different mountains. The second type (type B) includes entities that share the same name but belong to different categories. For instance, "肥遗" (Fei3 Yi2) in the *First Classic of Western Mountains* refers to a type of bird; while in the *First Classic of Southern Mountains*, it represents a type of snake.

For entities of type A, we assign a unique node ID to each of them. When building the knowledge graph, we handle these entities individually, while we process other entities in batches. For type B entities, we add category tags to them. This allows the computer program to identify the correct category of these entities based on the tag.

As for entity resolution, the phenomenon that entities in *Shan Jing* refer to the same object by different

names is primarily due to differences in full names and abbreviations. For example, in the *Second Classic of Southern Mountains*, the phrase "北望诸毗, 东望长右" (to the north it faces Zhu1 Pi2, to the east it faces Chang2 You4) uses the abbreviations "诸毗" and "长右" for "诸毗之山" (Mount Zhu1 Pi2) and "长右之山" (Mount Chang2 You4) respectively. During the entity resolution process, we manually expand these abbreviations into their full written forms.

4.4 Knowledge Base Visualization

We use the py2neo module in Python to connect to the Neo4j graph database. In this phase, we also add 26 chapter entities and 5 classic name entities to the knowledge base. As for the relationships between the classic name entities, the chapter entities, and the entities we previously extracted, we can refer to the schema in Figure 3. Therefore, the number of entities and relationships in our final knowledge graph is slightly higher than the statistical values given in Table 4 and Table 5. The total number of entities and relationships in the knowledge graph are 1432 and 3294 respectively. Figure 4 is a part of our final knowledge graph. It shows us the chapter structures in each section of *Shan Jing*. When you click on any chapter entity, the graph will display the mountain ranges contained in it. By further clicking on a particular mountain, you will be presented with the associated river and resource entities.

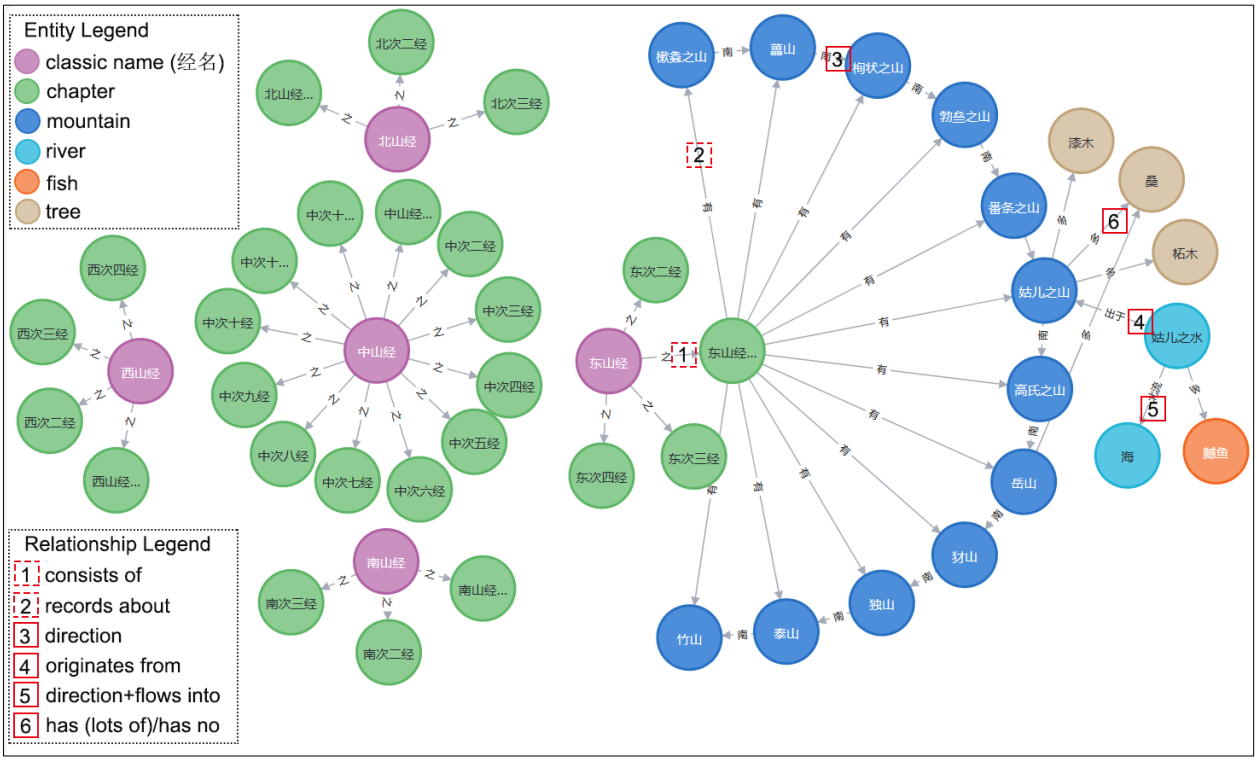

Figure 4: The knowledge graph about *Shan Jing*'s chapter structure.

5. Applications

Based on the knowledge base of *Shan Jing*, combined with the relevant works from other Chinese historical texts, building a Chinese historical ecological knowledge base is what we attempt to do in the future. To build the *Shan Jing* knowledge base, we systematically collect and classify the entities and relationships in *Shan Jing*, spending much time and effort to manually retrieve various related studies. The results provide the necessary database foundation for building a Chinese historical ecological knowledge base. As far as the study of *Shan Jing* or *Shanhai Jing* is concerned, the methods and the results of this paper can also provide help to related scholars. Here, the application value of this paper is only explained from the following three aspects: aiding in the collation of *Shan Jing* texts, assisting scholars to conduct quantitative studies, and providing a corpus base for *Shan Jing*'s monograph intelligent book.

5.1 Provide Assistance to the Collation of *Shan Jing* Text

Ancient Chinese is more likely to have some specific syntactic patterns than modern Chinese. As noted by Li (2020) during her utilization of regular expressions to analyze the poem titles in *Mao Shi Zhengyi* 毛诗正义 (*Annotations on the Book of Songs*), regular expressions assist in defining and summarizing general patterns. By examining and analyzing any exceptions that deviate from the pattern, they also aid in identifying errors in the text. Similarly, when applying regular expressions to match the text of *Shan Jing*, we discover their effectiveness in identifying textual errors within this book as well.

For example, when processing the *First Classic of Southern Mountains*, the regular expression fails to extract the mountain entity from the sentence "东三百里柢山" (Three hundred leagues further east is Mount Di3). Upon reviewing the annotations made

by previous scholars, we find that several scholars have suggested the inclusion of the character "曰" (called) before "柢山" (Yuan, 2016; Guo & Hao & Shen, 2015), which aligns with the rule defined in our regular expression. Similar problems arise in sentences such as "又西百五十里高山" (One hundred and fifty leagues further west is Mount Gao1), "（有兽焉）其名长右" (There is a kind of beast, its name is Chang You), "（有鸟焉）曰橐蜚" (There is a kind of bird, its name is Tuo2 Fei2), and so on.

5.2 Help Scholar to Conduct Quantitative Study

Traditional humanities research primarily relies on qualitative analysis of texts, although quantitative processing is occasionally included. Wu (2020) analyzed the mountains in *Shan Jing* that have no vegetation (无草木). He identified a total of 447 mountains, 95 of which were "without vegetation". The knowledge base of our paper can assist in such an analysis. Using the Cypher statement *MATCH(p1:Other classes{name: "草木"})<-[r:'无']-(p2)* in Neo4j, we can query mountains that fulfill the triplet (mountain, has no, vegetation). The result is depicted in Figure 5. The gray circle in the center of the image represents the entity "vegetation", while the dark blue entities scattered around it represent the mountains that are "without vegetation". The total number of 95 mountains matches the statistics We mentioned in his article.

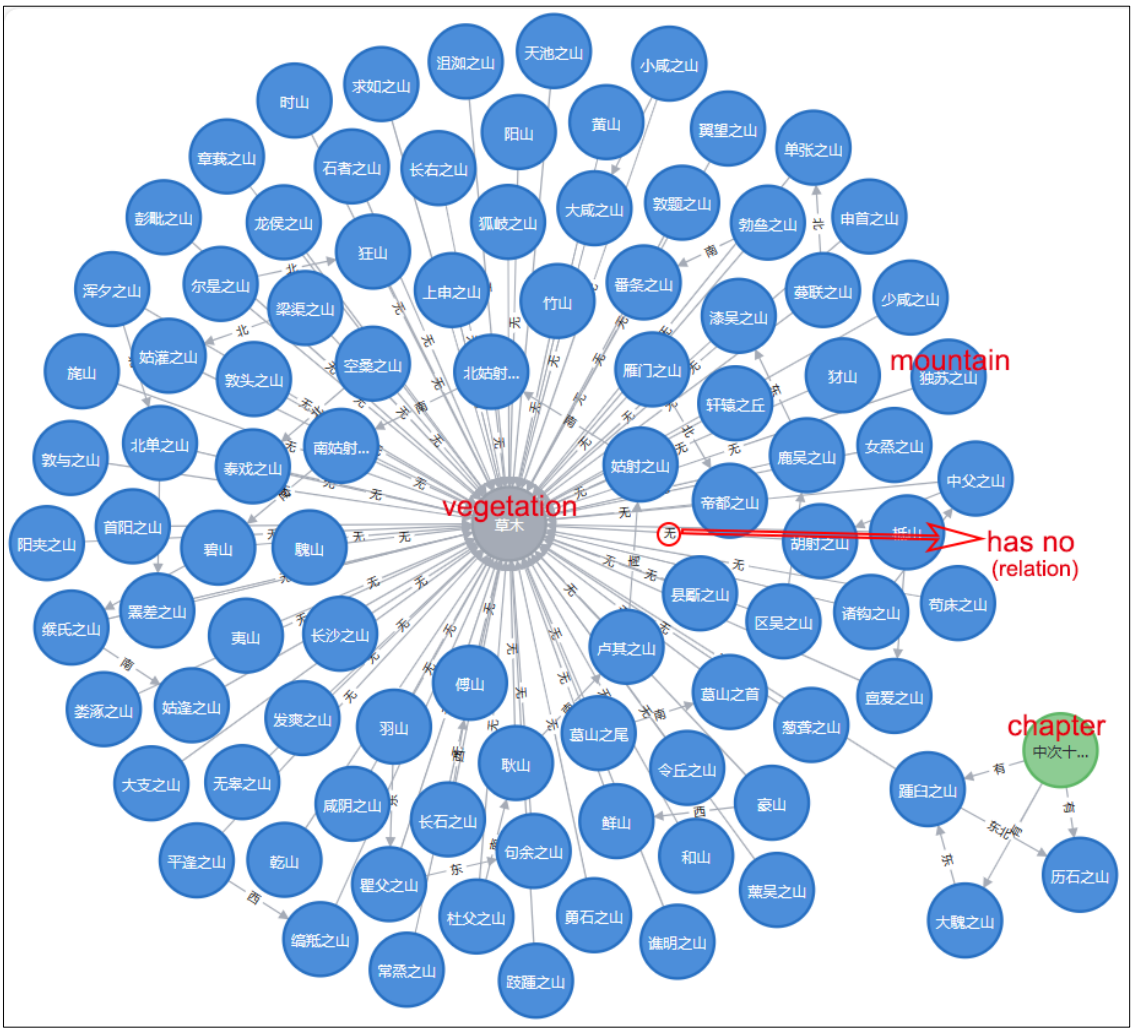

Figure 5: The knowledge graph about mountains without vegetation.

5.3 Provide Corpus Basis for *Shan Jing*'s Monograph Intelligent Book

Shanhai Jing is of great importance in ancient Chinese culture. When discussing topics of ancient Chinese mythology, geography, and ecosystems, it is impossible to exclude *Shanhai Jing* from the conversation.

Since Guo Pu's first annotation about *Shanhai Jing*, scholars from different dynasties have shown varying degrees of interest in this book, with many putting considerable effort into analyzing and explaining the entities mentioned in it. Table 6 presents a collection of monographs on *Shanhai Jing* written before the Republic of China. These works have served as important references for modern scholars who wish to engage in the study of *Shanhai Jing*. Examples include Yuan Ke (2016) and his *Shanhai Jing Full Translation*, Shen Haibo (2015) and his collated version of *Shanhai Jing*, and Guo Fu (2004) and his scholarly work *Attestment to Shanhai Jing*.

Dynasty	Author	Book Name
Jin	Guo Pu	*Annotations on Shanhai Jing* 山海经注
Jin	Guo Pu	*Panegyric on Illustrations of Shanhai Jing* 山海经图赞
Ming	Wang Chongqing	*Interpretations of Shanhai Jing* 山海经释义
Ming	Liu Huimeng	*Critiques of Shanhai Jing* 评山海经
Ming	Yang Sheng	*Supplementary Annotations on Shanhai Jing* 山海经补注
Ming	Zhu Quan	*Elaborate Words on Shanhai Jing* 山海经腴词
Qing	Wu Renchen	*Extensive Annotations on Shanhai Jing* 山海经广注
Qing	Wang Fu	*Surviving Texts of Shanhai Jing* 山海经存
Qing	Bi Yuan	*New Collation of Shanhai Jing* 山海经新校正
Qing	Hao Yixing	*Annotation and Commentary on Shanhai Jing* 山海经笺疏
Qing	Chen Fengheng	*Collected Commentaries on Shanhai Jing* 山海经汇说
Qing	Yu Yue	*Read Shanhaijing* 读山海经

Table 6: Monographs on the study of *Shanhai Jing* before the Republic of China

Since most of these monographs analyze both entities and their associated annotations, we can break down the scholars' research into individual annotations. For each entity, we collect and organize these annotations from different scholars and present them collectively to readers. This approach allows researchers to gain a comprehensive understanding of the entity efficiently, saving them time.

Figure 6 shows the interface designed for the intelligent book of *Shan Jing's* monographs. The leftmost column displays the chapter list of *Shan Jing*, while the center section presents a text fragment extracted from *Shan Jing*. In the upper right corner, a knowledge graph corresponding to the text in the center is displayed. When hovering the mouse cursor over a proper name in the text or an entity in the knowledge graph, such as "狌狌" (Sheng Sheng) in the image, all annotations by previous scholars on that entity are automatically displayed in the lower right corner. This approach increases the reader's convenience compared to the traditional method where researchers have to go through extensive studies for a comprehensive understanding about the entity. The use of intelligent books offers a solution that mitigates the criticism of academic fragmentation and disconnection often associated with traditional humanities research (Burdick et al., 2018). In this way, with the widespread dissemination of knowledge bases, accessibility to comprehensive information can be improved.

6. Conclusion

We report in this paper the construction of a knowledge base for *Shan Jing* as the first step toward recovering the ecological knowledge of China two millennia ago. The knowledge base covers 15 categories of entities and 6 types of relationships and provides both a foundation and a historical baseline for future constructions of China's historical and ecological Knowledge databases. This study also adds the dimension of the annotation of geographic and environmental information to ancient language processing. In addition to the extension to cover the entire *Shanhai Jing*, including the inclusion of the descriptive texts and later annotations associated with these entities. A challenging further development is to develop a feature system of descriptive terms and the integration of a GIS with historical layers.

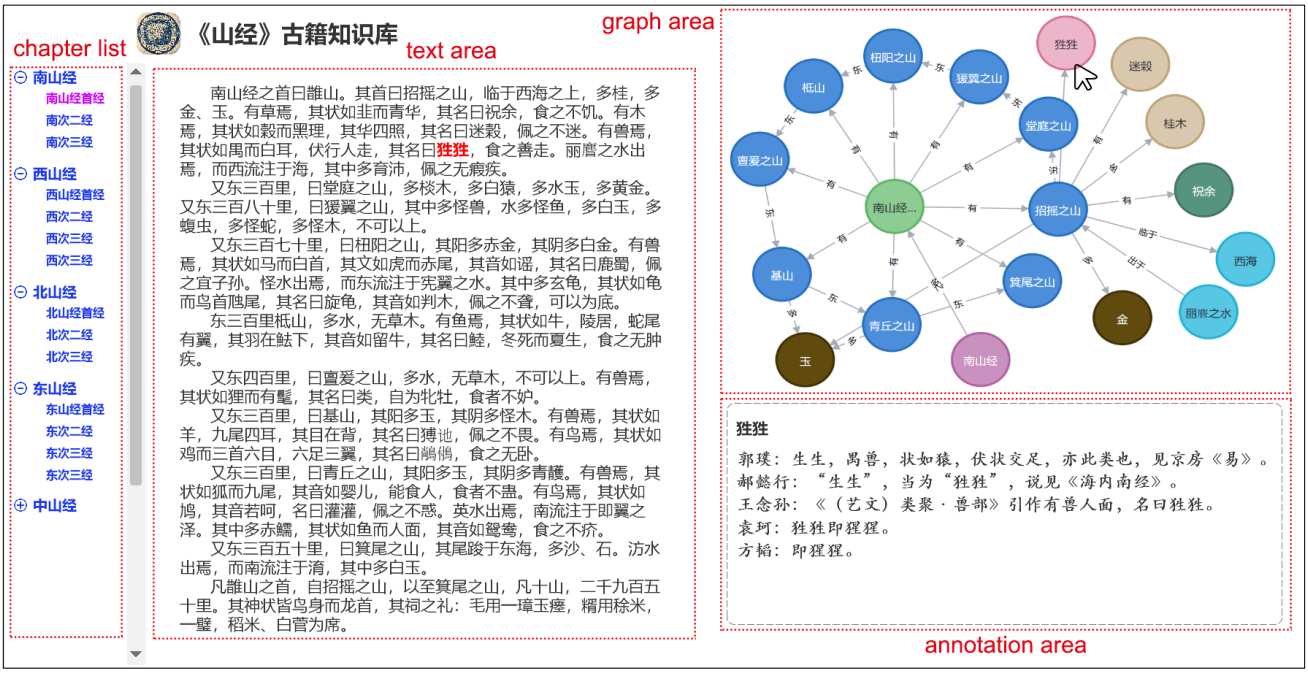

Figure 6: The interface designed for *Shan Jing* intelligent book.

7. Acknowledgments

The study reported is partly supported by the Major Laboratory Projects of the University of Chinese Academy of Social Sciences (Grant X20220112, Xinlan JIANG) and a PhD fellowship grant by the Hong Kong Polytechnic University (Ke LIANG). during the early stages of this project. We would like to thank Prof. Qi SU, Dr. Jinghang Gu, and Dr. Bo PENG for their advice at various stages of the project.

8. References

Birrell, A. (2000). *The classic of mountains and seas. Penguin*, London, 1st edition.

Burdick, A., Drucker, J., Lunenfeld, P. (2018). *Digital humanities: changing the rules of the game for knowledge innovation and sharing*. China Renmin University Press, Beijing, pp. 27.

Chen, L. S., (2012). *Investigation and discussion about the academic history of Shanhai Jing*. Peking University Press, Beijing.

Dorofeeva-Lichtmann, V. (2007). Mapless mapping: Did the maps of the Shan Hai Jing ever exist?. In *Graphics and Text in the Production of Technical Knowledge in China*, pp. 215-294.

Dorofeeva-Lichtmann, V. (1995). Conception of Terrestrial Organization in the Shan hai jing. *Bulletin de l'Ecole française d'Extrême-Orient*, 82, pp. 57-110.

Fu, X. Y. (2021). The "Proto-Ecological Narrative" in Shan Hai Jing. In Tang, W. S. (Translated), *Chinese Narratologies*, Beijing: Peking University Press, pp. 27-54.

Gagnon, A. C. & Berteaux, D. (2009). Integrating traditional ecological knowledge and ecological science: a question of scale. *Ecology and Society*, 14(2).

Guo, P., Hao, Y. X., Shen, H. B. (2015). Shan-hai-jing. Shanghai Classics Publishing House, Shanghai, 1st edition.

Han, Y. Y. (1933). Table of plants and animals in Shan-hai-jing. *Folklore*, pp.45-104.

Han, Y. Y. (1936). The God Names in Shan-hai-jing. *Pui Ching Middle School Library Journal*, 2, pp. 11-15.

Hsieh, S. K., Chang, S. M., Chang, C. H., Zhou, Y. S., and Huang, C. R. (2006). In *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*.

Huang, C. R., Lo, F. J., Chang, R. Y., and Chang, S. M. (2004). Reconstructing the ontology of the Tang dynasty: A pilot study of the Shakespearean-garden approach. In *Preceedings of the OntoLex 2004 Workshop*, Pages 43-49, Lisbon.

Jia, W. H. (2004). A study of the proper names in The Classic of Mountains and Seas. (Doctoral dissertation). Retrived from [China Academic Journals (CNKI)].

Li, L. F. (2020). The application of regular expressions to describe and analyze textual patterns in ancient Chinese texts: A case study of the poem titles in Maoshi Zhengyi. *Digital Humanities*, (2), pp. 75-89.

Liu, Z. D., (2006). Lost Scripture: The Classic of Mountains and Seas and the Ancient Chinese Worldview. The Commercial Press, Beijing.

McGregor, D. (2008). Linking traditional ecological knowledge and western science: aboringal perspective from the 2000 state of the lakes ecosystem conference. *The Canadian Journal of Native Studies* XXVIII(1), pp. 139-158.

Mustonen, T. (2019). Meaningful engagement and oral histories of the indigenous peoples of the north. *Nordia Geographical Publication*, 47(5), pp. 21-38.

Que, W. (1995). Historical geography in China. *Journal of Historical Geography*, 21(4), pp. 361.

Strassberg, R. E. (Ed.). (2018). A Chinese bestiary: Strange creatures from the guideways through mountains and seas. Univ of California Press, California, U. S., 1st edition.

Suh, K. H. (1993). A study of Shan Hai Ching: Ancient worldviews under transformation. (Doctoral dissertation). Harvard Univ, Cambridge, MA, USA. pp. 425-494.

Woo, F. J. (2002). Daoism and ecology: Ways within a cosmic landmark. *China Review International*, 9(1), pp. 112-118.

Wu, M. C., Lin, L. T., Qi, Y., Huang, S. Q., Wang, D. B., and Liu, L. (2023). Plant knowledge mining and organization construction in Pre-Qin classics from the perspective of digital humanities, *Library and Information Service*, 67(12), pp. 103-113.

Wu, X. D. (2020). The statistical analysis of 'wu cao mu', 'sha' and the related vocabularies in The Classic of Mountains, *Cultural Heritage*, 5, pp. 98-105.

Yuan, K. (1980). Appendix of Shan-hai-jing jiao-zhu. Shanghai Classics Publishing House, Shanghai, pp. 1-128.

Yuan, K. (2016). *Shan-hai-jing full translation*. Beijing United Publishing Corporation, Beijing.