International Journal of Knowledge and Language Processing
Volume 11, Number 2, **2020**

www.ijklp.org
KLP International ©2020 ISSN 2191-2734
pp.31–38

The Preliminary Study of the Construction of the Structured Corpus of the Pre-Qin Literature and Its Annotated Documents

Zhang Yanhua
School of International Education
Nanjing University of Science and Technology
No.200 Xiaolingwei Street, Xuanwu District, Nanjing 210094, China
zhangyh712@163.com

Received May 2020; revised July 2020

Abstract. *The structured corpus of the Pre-Qin literature and its annotated documents provide raw data and its format for Pre-Qin literature information processing. This paper discusses the main work on the digitalization of Pre-Qin literature and its annotated documents, then proposes that we should construct XML corpus as the final task of the above digitalization procedure. Based on this, we have designed and implemented the XML corpus of Pre-Qin literature and its annotated documents, which achieved the unified representation of meta-data and data of ancient books, and is helpful for alignment and search of Pre-Qin literature across different versions. We believe such corpora are the undisputed data sources and results for Pre-Qin literature information processing.*

Keywords: Ancient Books Digitalization, XML language, pre-Qin documents

1. **Introduction.** The information processing of the Pre-Qin documents refers to the comprehensive use of linguistic information processing techniques to conduct scientific analysis of the Pre-Qin documents and their annotations in terms of editions, collations, catalogues, annotations, proofs, forgeries, communication, codification, and retrieval [1], and then systematically discuss the generation, distribution, communication, utilization patterns, and language styles of the Pre-Qin documents and their annotations. The main experimental tasks of Pre-Qin documents information processing include automatic word segmentation, variant text discovery, part-of-speech word meaning tagging, syntactic analysis, etc. Objectively speaking, all the experiments of Pre-Qin documents information processing must be based on the digitalization of Pre-Qin documents and its annotations, and the experimental results must also be presented in the form of data structure or

literature informatization. Therefore, it is necessary to explore the digitization of the Pre-Qin documents and their corresponding annotated documents, especially the construction of a corpus of structured Pre-Qin annotated documents based on XML expression. These works will provide the raw corpus and data paradigms for the experiments on information processing of Pre-Qin documents, such as automatic participle, lexical semantic annotation, and syntactic analysis.

2. **Digitization of pre-Qin documents and their annotated documents.** The study of the digitization of the pre-Qin and its annotated documents falls under the category of digitization of ancient books. The so-called digitalization of ancient books can be roughly divided into two steps, electronic and informatization of ancient books. Electrification of ancient books refers to the use of modern information technology to transform ancient documents into the form of electronic media, preservation and dissemination through CD-ROMs, networks and other media, the main emphasis is for protection, storage and circulation of ancient books[2] (P47-48) ; Informatization of ancient books refers to the comprehensive use of information processing technology to reconstruct, retrieve, collate, compile and other in-depth computational analysis of ancient documents, the main emphasis is on the retrieval, compilation and processing of ancient books. Generally speaking, the broad definition of electronic antiquities should include the concept of informatization of antiquities, pure and narrow electronic antiquities cannot achieve the purpose of protecting the classics and transmitting them to today's use. This research paradigm has also been gradually eliminated by history. However, we take it in a narrow sense and divide the digitization of ancient books into two steps, namely, the electronic digitization of ancient books and the informatization of ancient books. Therefore, the electrification of ancient books is a necessary condition for the informatization of ancient books, and the informatization of ancient books is also the ultimate purpose of electrification of ancient books.

The electrification of ancient books is the future direction of ancient literature collation,[3] and is also a prerequisite for the digitization of ancient books and information processing of ancient documents. In general, the electrification of ancient books is conducive to maximize the preservation of good ancient resources, popularizing and spreading Chinese classical culture, and facilitating deeper informatization and information processing of ancient books, and therefore has attracted considerable attention from the academic community. The country's efforts to electrification antiquities date back to the early 1980s, and the large comprehensive database that has been developed includes the following.

(1) CD-ROM project on "China's Basic Antiquarian Library". The database was built by Mr. Junwen Liu of Peking University, with technical support from North Founder's Institute of Technology. From more than 100,000 kinds of ancient books, the database selects more than 10,000 kinds of codices from the Pre-Qin Dynasty to the Republic of China, divided into four sub-repositories: philosophical, historical, artistic, and comprehensive, 20 major categories, 100 details. The full set of 500 discs, approximately 2 billion words, provides search services in 3 areas: classification, entry and full text.

(2) The electronic version of Imperial collection of four. The electronic version of Imperial collection of four of Wen Yuan Palace contains 3,400 books of all generations, with 700 million Chinese characters, produced by Shandong Huiwen Technology Development Center and Hong Kong Dizhi Culture Publishing Co. The latter, while maintaining the authenticity of the original book, is indexed in its entirety, thus providing the reader with a quick and efficient means of retrieval, statistics, collation and editing.

(3) The National Science Dictionary (http://www.gxbd.com). The database was developed under the auspices of the Institute of Electronic Documentation of the Capital Normal University, and contains more than 3,800 ancient books with a total word count of more than 900 million words, and provides an online version for users to search remotely. At present, the database continues to expand at the rate of 200 million to 300 million words per year, which is of some typical significance for the study of information technology of ancient books in conjunction with the Internet.

(4) Project for the electronic dissemination of philosophy books (http://ctext.org). The Chinese Philosophy Book Electronic Project is one of the online open source classical literature sources. The project is a collection of classical research resources, including original dictionaries, contemporary research databases and internal dictionaries. Some of the original texts are available in both English and Chinese. There are also advanced search features such as an internal dictionary, a scanned version of the base text, and similar paragraph tips.

Other famous ancient book databases include: the database of ancient books of the Qin-Han dynasty, the northern and southern dynasties of Wei Jin, bamboo, silk and oracle books, the database of the Central Research Institute of Taiwan's Chinese Electronic Documents series, and the digitalization project of ancient books launched by the Superstar Digital Library. It should be said that these databases are the results of electronic research on ancient books, but many of them already provide high level information processing functions such as full-text citation, automatic search, similarity query, and so on, and are therefore the products of ancient book informatization.

Informatization of ancient books has been one of the focus of the academic community in recent years. As the original purely electronic ancient books are increasingly unable to meet the research needs of the academic community, many scholars have begun to carry out high-precision electronic book, information citation, ancient book collation and other in-depth information processing work. These efforts, which can be collectively referred to as the computerization of ancient books, include the following：

(1) Basic information processing. This part of the work is an extension of the electrification of ancient books, which mainly includes character set development, Rarely-used Chinese Characters supplementation and optical character recognition (OCR, Optical Character Recognition), etc. The aim is to restore the original appearance of ancient books as far as possible and provide the best electronic copy for subsequent information processing of ancient books. In contrast, this part of the work places more emphasis on the accuracy and fidelity of the literature than on purely electronic antiquities. Among them, the formulation of character set mainly discusses how to design a character set that meets

the needs of digitalization of ancient books under a unified framework. There is still a big gap between GB2312-80(6,763 simplified Chinese characters) and GBK standard (21,003 Chinese characters) and the number of characters used in ancient books. The CJK character set (70,195 Chinese characters) and Unicode implementation in ISO/IEC 10646 are the more suitable encoding choices for information processing in ancient books. The Rarely-used Chinese Characters supplementation means that, although the coverage of words in the ancient literature is nearly 100% as the character set continues to expand, it is neither economical nor practical to completely enumerate all the characters in the ancient literature, especially in the unearthed documents, due to the large number of variant and folk characters. Tanchun Qu has counted the folk characters in the relevant literature, and about 71.6% of the folk characters are extra-coded characters that need to be added. [4] Therefore, it is of great significance to design Rarely-used Chinese Characters supplementary tools for the use of archaeologists. We envision the ideal complementary tool for Rarely-used Chinese characters, where the user only needs to enter a number of character roots and the type of structure of the character form to complete the character creation process. The importance of Optical Character Recognition (OCR) for the documentation and information processing of ancient documents (including unearthed documents) is self-evident. Using optical character recognition technology, it is possible to convert ancient documents directly into machine-readable text with minimal manual proofreading. In addition, in addition to the technical requirements such as improving the recognition rate and adaptability, the OCR technology of ancient literature also needs to be studied to support traditional inline typesetting, automatic discovery of raw and folk characters and other related technologies.

(2) Multi-level citation search. This part of the work is the core work of ancient book informatization, the main process refers to correlating and structuring different versions of the same ancient book, followed by full-text indexing and annotation of related language knowledge, and writing corresponding search algorithms for research needs. According to our definition, the ancient documents obtained by the narrowly defined electronic translation of ancient books are merely text-coded sequences, lacking organic structural composition and data indexing, and therefore not conducive to subsequent information processing. We consider that the XML language achieves a unified representation of ancient book metadata and data, which facilitates the alignment and mapping of different versions of ancient books, and facilitates the user's information query and data processing for different needs, and has a variety of good online presentation effects, and should be used as a vehicle and target for structured information processing of ancient documents. For the digitization of the Pre-Qin and its annotated documents, the common XML language structure tags include: document name, author, annotator, chapter number, paragraph number, sentence number, sentence body, sentence annotation, "original sentence number", "word number", etc. [5] (p107-113). Based on this, we can conveniently implement full-text indexing and label the information elements in the literature with semantic syntactic information such as lexical markers, syntactic markers, named entity markers, and write retrieval and computational algorithms with different requirements according to the actual

needs, providing users with networked knowledge results.

(3) Collation and systematization of ancient documents. It has already been stated above that the information processing of Pre-Qin documents has been very purposeful since its birth, i.e. for classical bibliography, especially the digitization of classical documents and information processing services. Therefore, an important aspect of the Informatization of ancient books is to provide a new research perspective for the collation of ancient documents. This part of the work consists roughly of textual collation, Variant text discovery, exegesis, etc.

Text collation refers to the use of technical means to detect and correct any one-hundredth of a cent difference between electronic versions of the same document, and is therefore divided into two steps, "amendment" and "correction". Among them, the information processing of Pre-Qin documents can not only automatically discover different texts of the same ancient book, [6] but also provide rich data support for the final revision and erasure of the ancient documentarians. On this basis, we can also focus on the construction of a heterogeneous library. The main approach to the construction of a heterogeneous library is to store only one best electronic good copy for each literature, linking other versions of the heterogeneous text to the good copy on the basis of a full-text index of that text. This allows for both specialized querying of foreign content and the generation of different versions of electronic literature as needed. Clearly, it would be appropriate to use the XML language to build such an electronic database of good books. Annotation mining refers to the use of existing supervised or unsupervised machine learning methods to carry out machine learning on word knowledge, grammatical knowledge, semantic knowledge and even historical knowledge on the interpretation of annotations in the database of ancient documents, the results of which can be constructed into a detailed knowledge network of Pre-Qin documents, at the same time, it will also be of great benefit to information processing tasks such as automatic lexical division, lexical annotation, syntactic analysis, etc.

3. **Building a structured annotated literature corpus for XML expression.** Due to the limitations, the electronic texts of the Pre-Qin and its annotated documents are mainly collected from the Internet, but in fact they are only "half-finished" products that have been electronically converted from ancient books, with differences in character encoding, traditional and simplified characters, chapter order, and so on, and most of them have added vernacular punctuation and special format symbols to facilitate Internet transmission and reading. Therefore, the first step of our experiment was to unify the coding format of these electronic texts, unify the wording, unify the pre-processing of the chapter table of contents, and make the necessary errata and revisions according to the paper version.

These pre-processing steps will not be repeated here, but it should be pointed out clearly that the materials obtained from the pre-processing are still just textual sequences, lacking organic structure and elemental markings, and cannot yet be called "informatization of ancient books" or "digitization of ancient books", which is also not conducive to subsequent information processing. Based on this consideration, we first discuss how to construct a structured corpus of annotated literature based on XML language expression.

The full name of XML is Extensible Markup Language, or "Extensible Markup Language", which is designed to transfer and store data, and its nature is an intermediate, structured, generalized meta-language, so it is also suitable for storing and transferring ancient documents, as well as ancient documents information.

The ML record of ancient books is a new research paradigm that has emerged in recent years with the informatization of ancient books. Chuan Shan et al. (2007) applied XMLL and XML Schema to describe ancient book metadata, providing a basic scheme for recording ancient book metadata [7](p53-56); Qinxia Wu et al. (2009) used XML to structure the Oracle corpus [8](P5185-5188); Jihua Song et al. (2007) designed the corresponding annotation specification for the Sayings and Decoding Characters and carried out the full text annotation[9]; Xinxin Ma et al. (2013) introduced how to describe the Analects and its annotated literature based on XML, and how to organize the knowledge in multiple annotated literature to form a knowledge network [10]. These works show that the effects of using the XML language for antiquarian documentation and information processing are empirical. Specifically, the advantages are the following.

The XML language implements a unified representation of ancient metadata and data, categorizing and describing the ancient texts and themselves. For example, we can develop an XML Schema to represent the authorship, version information, classification index, and various information in the content of ancient documents. At the same time, the representation of the XML language has the generality, can be unified expression of arbitrary ancient literature, so as to facilitate the organization of a larger collection of literature, and also lay the foundation for comparison and calculation between different documents.

The XML language uses a structured tree format to describe data, making it easier for users to query information and extract data. As the corpus grows in size, or as the task deepens, the XML language can easily scale up or down the markup system to meet changing needs. In addition, corpus using XML language markup can be imported into relational databases for easy organization, indexing and retrieval of knowledge.

The XML language breaks the linear order of text and forms a knowledge network in a hyperlinked, fine-grained way to facilitate the retrieval, expression and calculation of the specified content. Text stored in the XML language is assigned a uniquely indexed hyperlink address, so it can be quickly jumped from one text to another, or the specified content can be extracted as much as the user wants. This is consistent with the multiple cross-relations between the Pre-Qin documents and the annotated documents, and also facilitates the mapping between the text and citations, as well as the annotations.

Of course, the use of XML language structured in the Pre-Qin and its annotated literature, its advantage also lies in the separation of data, display and operation, in line with the MVC software development framework; allows the same literature with a variety of display effects, easy for users to personalize the search, etc. From the practical point of view of the Pre-Qin and annotated literature, we have designed a set of annotation specifications to handle. In the scheme implementation, we use XMLSpy software to specify the XML

Schema pattern, with the program to automatically generate XML documents.

(1) For the original text of the pre-Qin documents, the following information can be described after the pre-processing and participle.

Metadata: information on the author of the document; information on the age of the document; information on the version of the document.

Data: Chapter information and numbering; paragraph numbering; sub-sentence numbering; sentence body; subtext; number of characters; number of words.

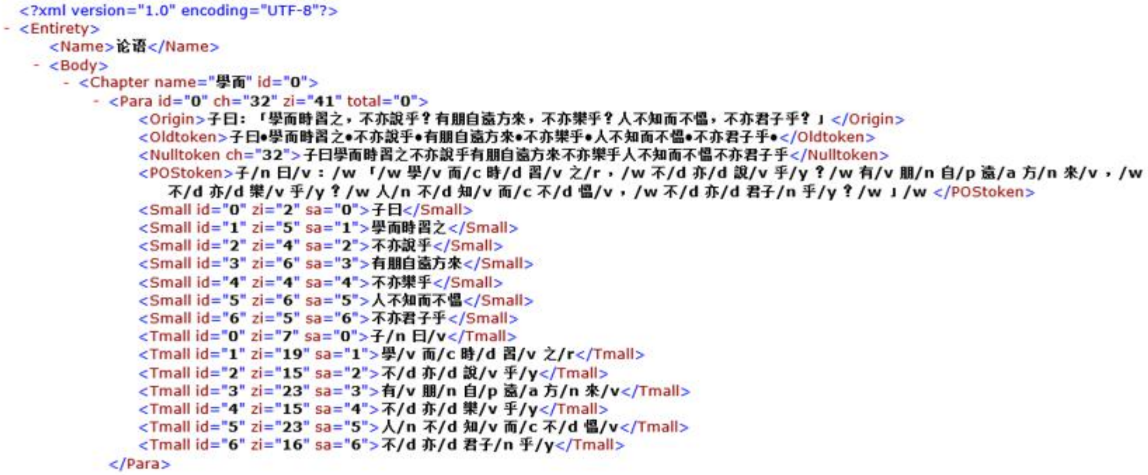
(2) For Pre-Qin documents annotations, after preprocessing, annotation alignment and word segmentation, the following information can be described:

Metadata: the author information of annotated documents; Announcing the chronological information of documents; Edition information of annotated documents;

Data: chapter information and number; Paragraph number; Clause number; The body of the sentence; Word segmentation text; Number of characters; Number of words; Citation number; Citation mapping text clause number; Note number; Citation number of annotation map; Number of the sparse text; Annotation number of sparse text mapping; Comment content; Annotation content maps text clauses and word information;

According to the actual needs, we can also expand the marking system, such as: the content of variant texts, syntactic analysis results, naming entity information, etc.; at the same time, we can also select only part of the marking information and elemental content for experiments, for example, in the automatic sentence break punctuation task of the Pre-Qin documents and notes, we do not need to mark the participle and notes information, only need to output the sentence content and number.

The following figure shows a relatively complete XML file (The Analects of Confucius - XueEr), the illustrations and code are generated by XMLSpy production.



```
  <?xml version="1.0" encoding="UTF-8"?>
- <Entirety>
     <Name>论语</Name>
   - <Body>
     - <Chapter name="學而" id="0">
       - <Para id="0" ch="32" zi="41" total="0">
            <Origin>子曰：「學而時習之，不亦說乎？有朋自遠方來，不亦樂乎？人不知而不慍，不亦君子乎？」</Origin>
            <Oldtoken>子曰•學而時習之•不亦說乎•有朋自遠方來•不亦樂乎•人不知而不慍•不亦君子乎•</Oldtoken>
            <Nulltoken ch="32">子曰學而時習之不亦說乎有朋自遠方來不亦樂乎人不知而不慍不亦君子乎</Nulltoken>
            <POStoken>子/n 曰/v ：/w 「/w 學/v 而/c 時/d 習/v 之/r ，/w 不/d 亦/d 說/v 乎/y ？/w 有/v 朋/n 自/p 遠/a 方/n 來/v ，/w
               不/d 亦/d 樂/v 乎/y ？/w 人/n 不/d 知/v 而/c 不/d 慍/v ，/w 不/d 亦/d 君子/n 乎/y ？/w 」/w </POStoken>
            <Small id="0" zi="2" sa="0">子曰</Small>
            <Small id="1" zi="5" sa="1">學而時習之</Small>
            <Small id="2" zi="4" sa="2">不亦說乎</Small>
            <Small id="3" zi="6" sa="3">有朋自遠方來</Small>
            <Small id="4" zi="4" sa="4">不亦樂乎</Small>
            <Small id="5" zi="6" sa="5">人不知而不慍</Small>
            <Small id="6" zi="5" sa="6">不亦君子乎</Small>
            <Tmall id="0" zi="7" sa="0">子/n 曰/v</Tmall>
            <Tmall id="1" zi="19" sa="1">學/v 而/c 時/d 習/v 之/r</Tmall>
            <Tmall id="2" zi="15" sa="2">不/d 亦/d 說/v 乎/y</Tmall>
            <Tmall id="3" zi="23" sa="3">有/v 朋/n 自/p 遠/a 方/n 來/v</Tmall>
            <Tmall id="4" zi="15" sa="4">不/d 亦/d 樂/v 乎/y</Tmall>
            <Tmall id="5" zi="23" sa="5">人/n 不/d 知/v 而/c 不/d 慍/v</Tmall>
            <Tmall id="6" zi="16" sa="6">不/d 亦/d 君子/n 乎/y</Tmall>
         </Para>
```

FIGURE 1. EXAMPLE.

4. **Conclusions.** At present, we have constructed a structured corpus of XML expressions of pre-Qin documents such as The Analects of Confucius and its annotations, and are continuously expanding it.

We investigate the digitization of pre-Qin documents and their corresponding annotations, and try to build a structured corpus of annotated documents based on XML expressions, in order to provide the original corpus and basic support for information processing of Pre-Qin documents. Next, we will continue to study the automatic sentence breaking and automatic punctuation of the Pre-Qin documents and their annotations, explore the lexical knowledge mining techniques based on the annotated documents, and conduct automatic word splitting experiments using the annotated documents, with a view to contributing to the speedy realization of information processing in the Pre-Qin documents.

Acknowledgment. This Work is supported by Program of National Social Science Foundation of China (Grant No. 15BYY096).

REFERENCES

[1] Yingde Guo, et al, Theory and methods of Chinese classical literature [M], Beijing: Beijing Normal University Press, Beijing, 2008.

[2] Yang Chen, The results and problems of digitizing ancient Chinese books [J], Publishing Science, 2003, (4).

[3] Ling Cao, Research on digitization of ancient agricultural books [D], Nanjing: Nanjing Agricultural University, 2006.

[4] Tanchun Qu, A study on the method of character description of common characters [D], Nanjing: Nanjing Normal University, 2015.

[5] Chuangxin Ma and Xiaohe Chen, The XML-based Knowledge Representation of the Analects of the Confucius and its Annotated Literature Aligned Corpus [J], Book Intelligence Knowledge, 2013, (1).

[6] Xiaohe Chen and Minxuan Feng.et al, Information processing in the Pre-Qin documents [M], Beijing: World Book Press, 2013.

[7] Chuan Shan and Chenguang Luo, A preliminary study of XML metadata for archival book recording[J], Library Work and Research, 2007, (6).

[8] Qinxia Wu and Yongge Liu, A study of XML/Schema oracle corpus-based corpus annotation [J], Science, Technology and Engineering, 2009, (17).

[9] Jihu Song, Ning Wang, and Jiajia Hu, A corpus-based approach to the construction of a digital "Sayings" research environment [J], Language and Text Application, 2007, (1).

[10] Chuangxin Ma and Xiaohe Chen, Structured knowledge representation of annotated literature based on ontology and XML [J], Library Journal, 2017, (8).