Metadata Design for Chinese Medicine Digital Library Using XML

Christopher C. Yang* and Winter W.M. Chan**

*: Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
**: Department of Computer Science and Information Systems
The University of Hong Kong

Abstract
Chinese medicine development has been growing rapidly and it begins to draw significant attention globally. In Hong Kong, particularly, the government has targeted Chinese medicine as one of its major future developments. A system for registering Chinese medicine practitioners is being set up and a Chinese medicine port for medicine manufacturing and research is being investigated. Although it has drawn significant attention, information about Chinese medicine is not yet widely available on the Internet and searching for Chinese medicine information is extremely difficult. In this paper, we are presenting the XML metadata design for Chinese medicine. The system encompasses the register server, information servers, search engine, and client programs developed to browse and search XML documents available at distributed Chinese medicine digital libraries.

1. Introduction
Chinese medicine has thousands of years of history, however, the practice of Chinese medicine is not as formally and stringently developed as that of western medicine. In Hong Kong, higher education in Chinese medicine was recently established as the government identified Chinese medicine as a niche and aim to make Hong Kong a world center of Chinese medicine. Registration system for Chinese medicine practitioner will soon be set up in the year of 2000. The industry is also enthusiastic to construct a Chinese medicine port to develop good manufacturing practice facilities and new biomedical products and equipment. As a result, an advanced and easy to access Chinese medicine information service on the Internet is desired.

As the World Wide Web becomes popular, the information available on the web is extensive. Information overloading is a serious problem and users find it extremely difficult to search for the information they need. Commercial search engines that use traditional indexing and keyword matching are not able to provide a satisfactory performance to retrieve documents that match the user interest. The poor performance is also due to the lack of semantics markup on the Web resources. The Web information often relies only on the Hypertext Markup Language (HTML) is only machine-readable but not machine-understandable.

To improve the performance of searching, filtering, and processing of Chinese medicine information on the World Wide Web (WWW), we propose to apply metadata for Chinese medicine using eXtensible Markup Language (XML). XML is developed by World Wide Web Consortium (W3C), which is vendor-neutral international industry consortium founded to develop common protocol specifications for the evolving WWW. While HTML only allows structural markup of Web documents, XML enables Web authors to provide semantic markup so that the meaning of each particular element can be identified. Authors are now able to annotate and deploy domain-specific Web documents. It is perfect for our application on Chinese medicine information systems on the WWW.

In this paper, we present the design of metadata for Chinese medicine. The Document Type Declaration (DTD) and example of XML are presented. We have also developed a Java client/server based system to search and browse Chinese medicine XML documents.

* Christopher C. Yang is an assistant professor in the Department of Systems Engineering and engineering Management at the Chinese University of Hong Kong starting from January of 2000. He was an assistant professor in the Department of Computer Science and Information Systems at the University of Hong Kong.
2. Chinese Medicine

In the Chinese culture, we believe in the harmony of nature between heaven, earth, and man. Similar to any other beings in the world, man is composed of five elements 五行. These elements are metal金, wood木, water水, fire火 and earth土; and they are in mutual production and mutual conquest相生相克. For example, fire conquers metal but is controlled by water.

Man constitutes a microcosm in the macrocosm. Such a microcosm is balanced by two opposite cosmological forces of natural: Yin 陰 and Yang 陽. Different parts of the body correspond to different Yin and Yang and the five elements. For instances, fire corresponds to heart, small intestine and tongue; water corresponds to kidney and ears; wood corresponds to gall-bladder and eyes; metal corresponds to large intestine; and earth corresponds to stomach, spleen and mouth.

The Chinese medical practice draws upon the Chinese cultural beliefs to interpret the states of illness. There are three fundamental causative factors in diseases recognized by the traditional Chinese medicine practitioners: Yang 陽, Yin 陰, and combination of both. Yang is the external-causing agent including climatic, infectious, and contagious. Yin is the internal dysfunction. Accidental and traumatic injuries are considered as partly Yin and partly Yang.

In the Chinese medicine theory, any disharmony of the equilibrium of Yin and Yang or disturbance of their flow will procure weakness or illness. As a result, the task of a Chinese medicine practitioner is to restore such imbalance of these two cosmological forces. Herbal medicine is one of the major treatments. Other treatments, such as surgery, acupuncture, moxibustion, massage, breathing exercise, diet, etc., may also apply when necessary. The purpose of these treatments is to enhance or reduce Yin or Yang in order to reach the equilibrium.


Emperor's Manual of Corporeal Medicine is considered as the Bible of traditional Chinese medicine, however, the author is uncertain. It includes two treaties, i) questions and answers about living matter 素問 (Suwen) and the vital axis 靈樞 (Lingshu). It discusses the interrelationship between human and nature, and the basic theories of pathology, diagnosis, prevention and treatment.

Pharmacopoeia of Heavenly Husbandman explains the theory of Yin and Yang, drug interactions, indications. It includes 365 items of medicine of mineral, plant and animal. The book is divided into three parts. The first part is the belief of having the efficacy of nourishing or prolonging life. The second part is generally of non-toxic medicines and their uses to restore the constitution of the patient. The third category is the most toxic medicines and side effects, and is used for combating diseases.

Treatise on Febrile Diseases was written by Zhang Ji 張机 around AD 150-219. Manual of the Pulses was written by Wang Shuhe 王叔和 around AD 265-317.

The Great Pharmacopoeia was written by Li Shizhen 李時珍 in the Ming dynasty. It is the most influential and most commonly referred pharmacopoeia of traditional Chinese medicine among Chinese. It describes plants, substance of animal origin, minerals and metals together with their medical properties and applications. About 1,800 kinds of medicinal plants are described, in which 300 items have not been cited in previous works.

2.1 Chinese Medicine Development in Hong Kong

Although Chinese medicine has a long history of development in China and Chinese herbs are popular in every Chinese family, the professionalism is not yet well recognized in Hong Kong. It was only since the last ten years that the government began to understand the importance of its development. Unlike the practice of western medicine in Hong Kong, there isn't any registration system for the Chinese medicine practitioner until a committee is constituted to study the set up of a registration system in 1997. Recently, due to the financial turmoil, the government is seeking for
new directions to strengthen the economic growth. The government has identified Chinese medicine as a niche and aim to make Hong Kong as a world center of Chinese medicine.

The Hong Kong government has started to develop Traditional Chinese Medicine (TCM) since 1989. The Working Party on Chinese Medicine was first set up to review the use and practice of TCM in August of 1989. Then, the Secretary for Health and Welfare appointed the Preparatory Committee on Chinese Medicine (PCCM) in March of 1995 following the working party's recommendation. In March 1997, the PCCM submitted a report to make recommendation on the regulation and development of the practice, use and trading of TCM. The Administration accepted the recommendation and began to set up a registration system, developed some formal education in Chinese medicine, and encouraged further researches and developments in Chinese medicine. By the year of 2000, registration of Chinese medicine practitioners will begin. In the past few years, many degree programs were set up in several tertiary institutions.

In October 1997, the Chief Executive stated in his Policy Address that Hong Kong must set new course to develop high value added industries and services as the world had entered a new era of increasing competition. Consequently, the Chief Executive's Commission on Innovation and Technology, which was appointed in March of 1998, stated in the first report (September 1998) that they envisaged Hong Kong to be a world center for the development of Chinese Medicine in health food and pharmaceuticals. In October of 1998, the Chief Executive affirmed that the Chinese Medicine industry would enhance the economic growth through innovation and technology.

In 1999, a consortium, which is formed by several leading property developers, proposes to build a Chinese medicine port to follow the construction of a cyberport, which is approved in early 1999. The project includes three phrases: 1) to build a biotechnology center to develop Chinese medicine, 2) to develop good manufacturing practice (GMP) facilities in order to turn Chinese medicine into internationally recognized drugs, and 3) to develop new biomedical products and equipment to cure various diseases.

To echo the vision of the HKSAR government, the Chinese Medicine Digital Library (CMDL) project is proposed to leverage the development of Chinese medicine in Hong Kong. Although many new educational programs and research centers are set up recently to promote the Chinese medicine, there are very limited resources for the general public, Chinese medicine practitioners, and researchers to retrieve Chinese medicine information in the public domain.

2.2 Chinese Medicine Digital Library (CMDL) Project

Digital library research has drawn significant attention since the Digital Libraries Initiative I started in 1994. Six projects were conducted, where the Illinois Digital Library project focused on federating Repositories and semantic interoperability. The Digital Libraries Initiative II started in 1999, where research has extended to the social aspect. Questions such as how to make the digital libraries useful and usable are addressed. Digital library research was also active in the Pacific Asia region for the past couple of years. The First Asian Digital Library Workshop was held in Hong Kong in 1998. Research such as indexing and searching on oriental language documents were presented. Other application researches such as Chinese museum and financial digital library were also discussed.

Chinese medicine has always been popular in Hong Kong and now is beginning to gain the attention from other countries. However, there is still limited legal control over herbal medicines for the purpose of trading, registration, dispensing, and purchasing. There is also a lack of resources in obtaining knowledge in Chinese medicine. As a result, it is difficult to ensure good quality of its practice and use. A Chinese medicine digital library, which adopts the advanced information technology and originates in Hong Kong, is desired. It will also be helpful to promote the good quality of future Chinese medicine development and the image of Hong Kong as the center of Chinese medicine.

There are a few websites providing Chinese medicine information; however, the available information is very limited and only simple HTML format for structural markup is used. For examples, the URL of the Chinese Medicine Information Center in Taiwan is [http://www.herb.com.tw](http://www.herb.com.tw) and their information of
Chinese Herbs can be found at [http://www.herb.com.tw/ENGLISH/HERB_EX/pidx.htm](http://www.herb.com.tw/ENGLISH/HERB_EX/pidx.htm). Similar websites can also be found in Mainland China, Canada, United States, and New Zealand. They are all in HTML format where interfaces in CGI are available for submitting query. The markups on the data are only machine-readable but not machine-understandable. Typical keyword matching are utilized for searching where poor precision and recall are always obtained. In order to improve the usability of the web information of Chinese medicine, semantic markup is necessary. As a result, the meaning of each particular element can be identified during searching.

3. Metadata for Chinese Medicine

Metadata means data about data. The World Wide Web Consortium (W3C) defines metadata as machine understandable information about Web resources or other things [8]. Currently, the Web has huge amount of data; however, it has no classification of information [7]. Without metadata, the information available on the Web is only machine readable and human understandable, but it is not machine understandable. The lack of classification of information makes it difficult to handle data effectively and retrieve information efficiently. Although there are search engines available on the Web, which uses the traditional indexing and keyword matching to assist in retrieving relevant information, the performance is still unsatisfactory. Users always complain on the poor precision and recall. In order to improve such situation, it is recommended to use metadata to properly classify and structure information presented on the Web.

In this paper, we focus on the metadata design of Chinese medicine using eXtensible Markup Language (XML).

3.1 XML and DTD

XML is a simplified subset of Standard Generalized Markup Language (SGML). It maintains the features of extensibility, structure, and validation of SGML but is specifically designed for transmitting structured data to Web applications. The extensibility of XML allows authors to define new tag names and attributes names by specifying their syntax and semantics. The structure functionality of XML allows complex documents to be composed of simpler documents, that means, documents may be containers for other documents with arbitrary nesting. Validation allows authors to specify the structure of document by referencing a description of its grammar.

A XML document has both physical and logical structures. Physically, a XML document consists of storage units called entities. An entity may refer to other entities to cause their inclusion in the document. Logically, a XML document is composed of declarations, elements, comments, character references, and processing instructions. The logical and physical structures must nest properly.

In the physical structure of XML documents, entities are storage units. The entities have content and are identified by name. The document entity (or "root") serves as the starting point of a XML document. Entities can be parsed or unparsed. The contents of parsed entity are referred as its replacement text, which is considered as an integral part of the document. The content of unparsed entity may or may not be text and each unparsed entity has an associated notation.

In the logical structure of XML documents, the document is composed of declarations, elements, comments, character references, and processing instructions, all of which are indicated by explicit markup. Each XML document contains one or more elements. The boundaries of elements are by start-tags and end-tags or by empty-element tag. The name that identifies an element is called generic identifier (GI), which consists of a set of attribute specifications. Each attribute specification has a name and a value.

XML provides a mechanism, called the document type declaration (DTD), to define constraints on the logical structure and to support the use of predefined storage units. An XML document is valid if it has an associated DTD and the document complies with the constraints expressed in the DTD. That means, the DTD specifies the grammar for a class of documents. In each XML document, the body follows the DTD. Parsers rely on the DTD to validate the corresponding XML document.

In defining a DTD, certain syntax must be followed. In declaring an element entity, "<ELEMENT element_name (expression)>" will be used where 'element_name' is the name of the element and the 'expression' is the expression that the tag can enclose. The expression can be data or tags with any
combinations. Table 1 shows some possible combinations. In declaring an attribute, "<!ATTLIST element_name attribute_name
data_type rule>" is used where 'element_name' is the element which the attribute should be attached to. The 'attribute_name' is the name of the attribute. The 'data_type' is the data type of the attribute value. The 'rule' is the constraint for the attribute value; its value can be 'implied', 'default', 'required', etc.

Table 1. Combination of tags and data in the expression of DTD element entity

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>#PCDATA</td>
<td>data that should be evaluated.</td>
</tr>
<tr>
<td>CDATA</td>
<td>data that should not be evaluated.</td>
</tr>
<tr>
<td>a</td>
<td>a or b</td>
</tr>
<tr>
<td>a, b</td>
<td>b follows a</td>
</tr>
<tr>
<td>a*</td>
<td>zero or more a</td>
</tr>
<tr>
<td>a+</td>
<td>One or more a</td>
</tr>
<tr>
<td>a?</td>
<td>A or nothing</td>
</tr>
</tbody>
</table>

3.2 Markup of Chinese Medicine
Chinese medicine that we use daily can be categorized into three types: (a) Chinese herbs (中草藥), (b) proprietary Chinese medicine (成藥) and (c) recipes (or prescription, 慣用藥方). Each Chinese medicine has its own characteristic as found in the classical literature. However, metadata is not provided in this classical literature; as a result, searching for information is a difficult task.

Herbs are the fundamental components of Chinese medicine. They can be used in their fresh or dry forms. Figure 1 (a) and (b) show the fresh and dry form of Chinese herbs, 五味子, respectively. Proprietary Chinese medicine is usually a combination of several kinds of herbs, which has undergone manufacturing processes and appears in the form of tablets or capsules. Recipes are combination of different herbs prescribed by a Chinese medicine practitioner to achieve specific purpose of healing according to the identified symptoms, patient's age, gender, and health status, etc. Users are recommended to follow the instruction to prepare the prescribed herbs.

Figure 1. (a) fresh 五味子, (b) dry 五味子

3.3 Classifications and Properties of Chinese Medicine
In order to design a metadata set for the Chinese medicine, attributes of the medicines are first identified. Attributes of herbs, proprietary Chinese medicine, recipes are listed in the Table 2 to 4 and Figure 2. The listed attributes are based on the discussion with the local Chinese medicine practitioners, where Chinese medicines are usually organized with these attributes for practicing and teaching.
Figure 2. An overview of the Chinese medicine structure

Table 2. Attributes of Herbs

<table>
<thead>
<tr>
<th>Name (名稱)</th>
<th>Function (功能)</th>
<th>Used part of the herbs (藥用)</th>
<th>Taste (味)</th>
<th>Precautions (注意)</th>
<th>Related information (相關資訊)</th>
<th>Classification of the herbs (分類)</th>
<th>Appearance (圖片)</th>
<th>Characteristics of the herb (性質)</th>
<th>Chemical ingredients (化學成份)</th>
<th>Serving method (服用方法)</th>
</tr>
</thead>
</table>

Table 3. Attributes of proprietary Chinese medicine

<table>
<thead>
<tr>
<th>Name (名稱)</th>
<th>Function (功能)</th>
<th>Chemical ingredients (化學成份)</th>
<th>Serving method (服用方法)</th>
<th>Manufacturing method (製法)</th>
<th>Ingredients (成份)</th>
<th>Image (圖片)</th>
<th>Taste (味)</th>
<th>Precautions (注意)</th>
<th>Related information (相關資訊)</th>
<th>Manufacturer (製造商)</th>
</tr>
</thead>
</table>
Table 4. Attributes of recipes：

<table>
<thead>
<tr>
<th>Name (名稱)</th>
<th>Function (功能)</th>
<th>Chemical ingredients (化學成份)</th>
<th>Serving method (服用方法)</th>
<th>Ingredients (成份)</th>
<th>Target Figure (體格)</th>
<th>Target Interior (寒熱底)</th>
<th>Image (圖片)</th>
<th>Taste (味)</th>
<th>Precautions of using it (注意)</th>
<th>Related information (相關資訊)</th>
<th>Target Age (年齡)</th>
<th>Target Gender (性別)</th>
</tr>
</thead>
</table>

Figure 2 shows the acyclic labeled graph representing the overview of Chinese medicine structure. Ellipse is an indication of a class or an object. The ellipse on the first level corresponds to the document entity. The ellipses on the second level represent the three types of Chinese medicine. The ellipses on the third level correspond to the instances of each type of Chinese medicine. Each instance has its own attributes. The corresponding attributes and attributes' values are represented by arcs and rectangular boxes, respectively.

The DTD of Chinese Medicine XML document is divided into two parts. The first part is the classification of Chinese medicine and the second part is the properties of the three types of Chinese medicine. The DTDs are listed in Figure 3. The first line of declaration specifies the classification of Chinese medicine. It is enclosed by the tag . Within the tag , it can be the tags of , or . The last 3 lines of declarations specify the property set for each type of Chinese medicine. Given the property set, the declarations of all properties are given in Figure 4.

```xml
<!ELEMENT _中藥(_,中草藥_,_慣用藥方_,_成藥_)>  
<!ELEMENT _中草藥(_,名稱_,_圖片_,_分類_,_功能_,_藥用_,_性質_,_味_,_化學成份_,_注意_,_服用方法_,_相關資訊_)>  
<!ELEMENT _慣用藥方(_,名稱_,_圖片_,_年齡_,_體格_,_性別_,_寒熱底_,_功能_,_味_,_化學成份_,_注意_,_服用方法_,_相關資訊_,_成份_)>  
<!ELEMENT _成藥(_,名稱_,_圖片_,_功能_,_味_,_化學成份_,_注意_,_服用方法_,_相關資訊_,_製法_,_製造商_,_成份_)>  
```

Figure 3. The Document Type Declaration (DTD) of a Chinese Medicine XML document.

```xml
<!ELEMENT _名稱(_,中文名稱?,_英文名稱?)>  
<!ELEMENT _中文名稱 (#PCDATA)>  
<!ELEMENT _英文名稱 (#PCDATA)>  
<!ELEMENT _圖片 EMPTY>  
<!ATTLIST _圖片 SRC CDATA #REQUIRED>  
<!ELEMENT _分類 (#PCDATA)>  
<!ELEMENT _功能 (#PCDATA)>  
<!ELEMENT _藥用 (#PCDATA)>  
<!ELEMENT _性質 (#PCDATA)>  
<!ELEMENT _味 (辛/酸/苦/淡/甘/膩)[]>  
<!ELEMENT _化學成份 (#PCDATA)>  
<!ELEMENT _注意 (#PCDATA)>  
<!ELEMENT _服用方法 (#PCDATA)>  
```
Figure 4. Declaration of the Chinese Medicine properties.

The XML document of Chinese medicine is required to follow the specification of the logical structure described by the DTD. Otherwise, the XML document will not be validated and parsed by the parser.

**Example**

An example of Chinese Medicine XML document is shown in Figure 5.

```xml
<?xml version='1.0' encoding='Big5'?>
<!DOCTYPE _ SYSTEM "rdfxmltcm.dtd">
<_中草藥 xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
  <_中草藥>
    <rdf:RDF>
      <rdf:Description about="翠雲草">
        <_名稱>
          <_中文名稱>翠雲草</_中文名稱>
          <_英文名稱>Selaginella uncinata</_英文名稱>
        </_名稱>
        <_圖片 SRC='http://www.tcm.org/Selaginella uncinata.jpg' />
        <_分類>卷柏科</_分類>
        <_功能>清熱利濕，止血止咳。</_功能>
        <_藥用>全草。</_藥用>
        <_性質>性寒。</_性質>
        <_味>苦</_味>
      </rdf:Description>
    </rdf:RDF>
  </_中草藥>
</_中草藥>
```

Figure 5. A XML document of the Chinese Medicine.

**4. Applications of Chinese Medicine Metadata**

Given the metadata of Chinese medicine, applications such as searching, filtering, and processing will become more effective and efficient. Metadata provides the semantics of the content. Users can specify their specific interest of search and specific property values, searching can be processed without confusion with other properties.

A system, which includes a XML browser and a search engine using a client/server approach, is developed for application on the Internet. The register server and information servers are distributed at different sites. The information servers store the Chinese medicine information.
For examples, there could be a Chinese information server to store the information of Chinese herbs, and a Chinese information server to store the proprietary Chinese medicine and so on. These Chinese information servers are also known as the Chinese medicine digital libraries. The client program connecting to the register server through Remote Method Invocation (RMI) is able to obtain a list of information servers. Users can make queries to the list of information servers through the client program. The distributed digital libraries share the load of the huge amount of information. The register server is able to monitor the status and accessibility of the available digital libraries.

Given the selected information servers, the client program will create a customized interface based on the DTD of the XML files for users to submit their queries for searching. Figure 6 shows an example of a client interface providing fields for users to submit their queries. Figure 7 shows an output interface, which displays the results returned from all information servers.

The customized XML browser is flexible to show any XML documents. The layout of the document depends on the document structure. Users can click the arrow button to traverse the searching result. There may have a list of related links in documents. Figure 8 shows the representation of related links. Users can also follow the links to browse other XML files that provide Chinese medicine information. Once the links are clicked, a new XML browser will pop up showing the selected documents.

![Figure 6. Example of client interface for information searching.](image)

![Figure 7. Customized XML browser](image)
5. Conclusion
As Chinese medicine begins to draw attention, especially in Hong Kong, a Chinese medicine information system on the Web is desired. Using XML to provide the semantic markup, metadata is added to the Web documents of Chinese medicine. Given the metadata on Chinese medicine, the stored data become machine-understandable. Such machine-understandable data makes itself more usable, and hence, more manipulations of the data become possible. A system, which includes the information servers, register server, and client programs, is developed to enhance the searching, filtering and processing of XML documents. Using such a system, users are easier to retrieve the Chinese herbs, Chinese proprietary medicine, and recipes of their needs. Users can submit query using one of the Chinese medicine properties through different information servers and obtain the result on a tailor made graphical user interface.

References
9 http://www.w3.org/Metadata/