• 医学信息资源管理与利用 •

基于本体论构建中医古籍知识库的探索*

孙海舒 符永驰 张华敏 张金玲

（中国中医科学院中医药信息研究所 北京 100700） （中国中医科学院针灸研究所 北京 100700）

〔摘要〕 简要介绍本体、知识模型等基本概念和基于本体论构建中医古籍知识库的建库目标，阐明中医古籍知识表达相关理论，从规范控制、构建原则、本体构建工具、系统架构、术语规范化、构建本体模型等方面具体论述基于本体的中医古籍知识库的构建。

〔关键词〕 本体；中医古籍；知识表达；知识库

Exploration of Constructing Knowledge Base of Ancient Chinese Medicine Books Based on Ontology *SUN Hai – shu, FU Yong – chi, ZHANG Hua – min, Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China; ZHANG Jin – ling, Institute of Acupuncture and Moxibustion, China Academy of Chinese Medical Sciences, Beijing 100700, China*

〔**Abstract**〕 The paper briefly introduces the basic concepts of ontology, knowledge model and the aims of constructing knowledge base for ancient Chinese medicine books based on ontology, clarifies the theories related to ancient Chinese medicine books knowledge representation, concretely discusses the construction of ancient Chinese medicine books knowledge base based on ontology from the following aspects: standard control, construction principle, tools for ontology construction, system framework, glossary standardization, constructing ontology model, etc.

〔**Keywords**〕 Ontology; Ancient Chinese medicine books; Knowledge representation; Knowledge base

中医古籍是现代图书馆文献体系的重要组成部分之一，其信息必须实现数字化和网络化。目前针对中医古籍元数据已经制订了一套符合自身规律的著录规则，其中包括资源形式、题名、主要责任者、其他责任者、出版项、附注说明、相关文献、主题词、语种、时空范围、古籍标识、馆藏信息、版本、外观形态、收藏历史的元数据项目，涵盖了绝大多数中医古籍著录的需求。在此基础上可将古籍元数据以计算机网络语言的形式进行描述，实现古籍信息的数字化。本文的目的就是希望在元数据的基础上，用基于本体的方法构建古籍多媒体知识库，用各种概念明确其内容，从而更好地实现古籍信息的管理、检索和共享。

1 基本概念与建库目标

1.1 本体

本体（Ontology）的概念最初起源于哲学领域，它在哲学中的定义为“对世界上客观存在物的系统的描述”。在信息系统、知识系统等领域，越来越多的人研究本体，并给出了许多不同的定义。

〔收稿日期〕 2010 – 12 – 17

〔作者简介〕 孙海舒，助理研究员，发表论文 6 篇。

〔基金项目〕 中国中医科学院自主选题“中医专家多媒体资源管理平台构建及方法示范研究”。

其中最著名并被引用得最为广泛的定义是由 Gruber 提出的："本体是概念化的明确的规范说明"[1-2]。

1.2 知识模型

知识模型，以本体为基本单位，突破以往图书馆仅仅以物理形态为文献管理单位的旧观念。同时从基本的知识维度出发建立每类知识关联的模型，作为联结本体的"语法规则"，将各种简单的知识模型组合成更复杂更专业的知识模型，直至最后建立其全部知识体系[3]。

1.3 建库目标

将"本体"引入古籍数字资源中来，可以使散见于各书或一书各篇、各卷间的某种特定信息集中并使其具有某种关联性，从而达到重整内容资源的目的，实现知识发现的功能。这种本体的建立不限于特定名称的抽取、标引和定义，同时由于人文历史领域知识有着较强的时空依赖性、主观性、不确定性、模糊性和争议性，因此更为重要的是要建立适用于具备某种特性或关系的不确定因素的本体。时间模糊查询即查询特定的时间段落或周期中数据，例如检索早于某年、晚于某年或某年至某年之间的数据或是提取多年以来某个季节的数据；地理信息的模糊查询即检索同属于某些行政区划或地形地域的数据，例如检索属于唐代医官范围内的碑铭资料；分类模糊查询即检索同属于某些特定类别的事物，例如检索属于百合科的植物或是查找属于史部传记类的古籍；方剂配伍关系模糊查询即检索属于某些特定证候、特定疾病、特定配伍关系的数据，例如检索万历时期名医吴元溟的著作等等类似的模糊查询要求还有很多。拟建立药名规范模型、古今药名沿革模型，以满足药名信息模糊查询要求；拟建立年表模型，以满足时代模糊查询要求；拟建立方剂配伍模型，以满足分类模糊查询要求；拟建立配伍关系模型（相须关系、相恶关系、相使关系、相杀关系等等）、中药"籍贯"（道地药材）模型，以满足中药、方剂等的模糊查询要求；拟建立穴位模型、穴位组配模型，以适应针对针灸的模糊查询要求[4-7]。

2 中医古籍知识表达

2.1 知识表达的涵义和意义

所谓知识表达，是指将知识通过某种数据结构结合到计算机系统中的程序设计过程，是知识的模型化和形式化，目的是便于知识在计算机中的存储、检索、使用和修改。构建中医古籍领域本体的目的是为了根据中医古籍知识库中的现有知识，实现中医临床、科研、教学知识共享、应用等功能。例如，根据提供古籍中病例的症状，根据系统中的知识分析出证候、做出判断，进而推导出针对现代临床病例相应的治法，临床医生根据辩证施治，提出个体化的诊疗方案或干预措施。

2.2 数字时代的中医古籍知识表达

据统计，《全国中医图书联合目录》收载中医药图书 12 124 种，其中古籍文献 8 000 余种。数字时代的到来，信息科学和网络技术的迅猛发展，已为更广泛的知识组织工作提供了良好的条件。目前，真实意义上的知识组织工作已在中医药古籍研究领域展开。中医古籍文献的知识表达工作，主要分 3 个步骤。一是实现古籍载体形式的变更，将纸张载体文献变成数字化古籍，为保存、整理和利用古籍资源奠定基础。二是确定中医古籍的知识表示方法，通过知识解析和标注对数字化古籍进行深度加工，以促进知识的科学组织和有效利用。三是构建知识库，利用人工智能中的机器学习、知识处理和神经网络等方法，实现知识因子的有序化和知识关联的网络化，从知识库中挖掘有用信息，发现知识，为用户提供有效的服务。

2.3 中医古籍的标引与知识表达

知识表达是研究知识从自然记载形式过渡到适合计算机处理的表示形式，在此基础上实现对知识的处理。中医古籍知识解析标引过程，是以古代文献的知识内容与结构为基础，结合信息学理论及计算机的知识表示方法，采用 XML 技术将中医古文献结构化的工作过程。在古籍文献解析与标注过程

中，要求专家基于对原文献结构层次的分析，将其分解成知识体、知识元、语义成分等逻辑层次分明的结构化文档，并根据其对知识内容的理解，提取知识元、知识体的概念。在研究过程中逐渐确立了划分知识层次、概念提取的原则，规定了药、证候、附方、语义成分、注文、眉批、歌赋、文中小字、校勘记等内容的具体解析标注方法。经过多年实践以知识为核心的中医古籍知识表达和知识解析体系逐渐从雏形走向成熟。这不仅为数字时代的文献整理研究开辟了一条新路，也为基于知识的中医古籍知识表达方法提供了一个成功的案例[8]。

3 中医古籍数字资源知识库的建设

3.1 基本理念

3.1.1 基本模式 国际图联 1997 年研究报告《书目记录的功能需求》中指出中医古籍数字资源知识库基本模式是从文献实体角度提出的著作、品种、版本和复本 4 个层次类型、相关属性及其互相之间的关系。其特点在于强调对信息的分层级管理，实现同一著作不同品种间的参见、同一品种不同版本间的参见等关联。这是一个共性的概念，同样适合中医类古籍。

3.1.2 规范控制 规范控制指为了确保文献标目的一致性以对标目实现统一管理的手段，包括以下内容：规范标目、参见标目、规范标目与相关标目间的参照关系以及选取标目及确定其参照关系的依据。规范控制理论广泛应用于纸本检索工具中并取得了很大成功。值得注意的是规范控制分为两个方面，即合并的规范控制和区分的规范控制，两方面结合起来才能得到最佳结果。合并的规范控制是将相同所指的不同关键词合并为一个款目，选择其中一个关键词作为规范标目，其他的作参见标目[9]。增加知识维度而使知识的扩展性升级、规范知识标引而提高检索效率、引入计量关系而强化知识关联的科学性是古典目录学与现代信息管理理论互相贯通的要点，也正是古籍数字资源知识库建设的基本理念。

3.1.3 构建原则 本文拟构建的中医领域本体基本符合 Gruber 于 1995 年提出的以下 5 条规则：（1）明确性和客观性。使用自然语言对术语给出明确、客观的语义定义。（2）完整性。给出的定义是完整的，能表达特定术语含义。（3）一致性。知识推理产生的结论与术语本身的含义不会产生矛盾。（4）最大单项可扩展性。向本体中添加通用或专用的术语时，通常无需修改已有内容。（5）最少约束。尽可能减少对建模对象列出约束限定条件。

目前关于中医本体的研究尚处于起步阶段，本文拟引用包含飞教授“中医顶层本体构建应用”的研究成果。构建中医本体是在语义层次发掘中医知识的基础，中医顶层本体不仅为中医本体的构建提供了框架，而且有利于实现中医本体同其他领域本体之间的整合，是构建完整中医本体的基础。鉴于知识的全球共享性，包含飞教授指出中医顶层本体必须包括：一般科学的概念接口、一般生物医学的概念接口、中医的最高层的抽象概念[10]。

3.2 构建方法

3.2.1 本体构建工具 根据惯例，每一个本体工程的开发项目应该参照 IEEE1074 – 1995 标准（软件开发生命周期法），应用一套相应的本体论构建方法。目前较常用的本体构建方法有 7 种：TOVE 法、METHONTOLOGY 法、骨架法、KACTUS 工程法、SENSUS 法、IDEF5 法、7 步法。根据现有研究成果，上述 7 种方法均允许在系统间进行互相操作，均提供知识共享和复用的机制，按成熟度依次为 7 步法、METHONTOLOGY 法、IDEF5 法、TOVE 法、骨架法、SENSUS 法、KACTUS 工程法。

其中 7 步法是由美国斯坦福大学医学院开发，主要用于领域本体的构建，它相对的本体编辑工具是 Protégé。Protégé 是在 Java 环境下开发出来的，与其他系统相比其优势在于：具有图形化的用户界面，能让用户进行可视化的编辑；支持 Unicode 字符集输入；可免费下载系统安装软件与插件；支持 DAML + OIL 以及 W3C 推出的 OWL，可以用 RDF，RDFS，OWL 等本体表示语言在系统外对本体进行编辑和修改，因此吸引了众多的使用者。其中本体视图插件可让用户直接看到本体的图形化表示，还

为用户直接配置了 OWL 插件，供本体以 OWL 格式存储，并在系统外进行其 OWL 文件的编辑。目前国内一些相关研究，例如上海中医药大学中医药信息化－标准化研究室的本体构建研究、南京理工大学“基于本体的中医学脾胃病知识库的构建”等主要采用 Protégé 工具，中医领域存在大量的非结构化知识、中医古籍知识的提取是包含于中医领域的，但同时有自身的特殊性。本研究拟构建的中医古籍领域本体，较多地参考其方法和流程，并结合中医古籍内容自身特点进行。

3.2.2 *系统架构* 建立中医古籍知识本体，除了需要本体构建工具，还需要本体构建的系统框架。本文所探讨的中医古籍本体研究尚处于理论研究与论证阶段，根据国内相关研究成果，拟采用 Jena 框架结构。Jena 是惠普（HP）公司开发的一个基于 Java 的开放源代码语义网工具包，为解析 RDF，RDFS 和 OWL 本体提供了一个编程环境及一个基于规则的推理引擎。Jena 有以下几项主要功能：（1）RDF API。可将 RDF 模型视为一组 RDFstatements 集合。（2）RDQL 查询语言。对 RDF 数据的查询语言，可以伴随关系数据库存储一起使用以实现查询优化。（3）推理子系统。包括基于 RDFS，OWL 等规则集的推理，也可自己建立规则。（4）内存存储和永久性存储。Jena 提供了基于内存暂时存储的 RDF 模型方法，目前支持 MySQL，Oracle 和 PostgreSQL 的数据存储。（5）本体子系统。Jena 对 OWL，DAML + OIL 和 RDFS 提供不同的接口支持。基于上述优点本研究拟采用 Jena 框架结构。本文重点关注基于本体构建中医古籍知识库，因此关于 Jena 语义网框架中的接口结构、推理子系统等暂不做细节讨论。

3.2.3 *中医古籍术语规范化* 规范化的医学术语集是医学领域的知识本体。医疗信息在人与计算机之间、计算机与计算机之间精确地识别与传递，则有赖于规范化的医学术语集为其提供底层支撑。从知识本体的定义来看，规范化的医学术语集就是属于医学领域的知识本体。国际著名的系统化临床医学术语集（Systemized Nomenclature of Medicine, Clinical Terms, SNOMED CT），是医学领域成功利用知识本体的研究成果。目前中医领域的中医药一体化语言系统（Traditional Chinese Medical Language System, TCMLS），已经是一个比较成熟的语言系统，本研究在对中医古籍术语进行规范化的同时采用的原数据主要基于此。但是中医古籍术语不完全等同于中医临床术语，因此元数据的选择还参考了《我国数字图书馆标准规范建设——古籍描述元数据规范》，并且力图将两者结合应用，建立中医古籍术语规范，使其针对中医古籍的知识提取具有可操作性。具体方法如下：（1）确定专业领域和范围。构建本体时必须明确目标，本研究以中医古籍作为特定的研究领域，以基于中医古籍领域本体的知识查询和检索为应用目的，探讨基于本体论的领域知识组织方法。（2）复用现有本体的可能性。一部分基于已经规范化的中医临床术语，其余部分尚无可用的现成本体。

3.2.4 *利用 Protégé 构建中医古籍本体模型* 关系代表了领域中概念的交互作用[11]。基本的关系共有 4 种，见表 1。

表 1 基本关系

关系名	关系描述
Part－of	表达概念之间部分与整体的关系
Kind－of	表达概念之间的继承关系，类似于面向对象中的父类与子类之间的关系
Instance－of	表达概念的实例与概念的关系
Attribute－of	表达某个概念是另一个概念的属性，如“价格”是“桌子”的一个属性

类是本体中最基本的组织单元，代表了一类具有共性的实例对象，是一种层次结构的组织形式。子类可继承父类的抽象特性，代表比父类更具体的概念。如“证候”是证候本体的最高层次类，而八纲证候、脏腑证候等是其子类，它们会继承证候的所有特性。定义类和类层次的基本方法是：从概念集合中选取具有独立存在性的对象概念（不是描述这些对象性质的概念），作为本体类层次结构中的锚点。通过判断某个类的实例是否也是另外类的实例来判断两个类的层次关系。本文以中医古籍中中医诊断部分的证候数据为例来说明具体构建方法。证候本体设计过程中，在保留中医辨证特色的原则下，将不同的辨证方法下形成的证候，如病因证

候、八纲证候、气血津液证候、脏腑证候、经络证候、六经证候、卫气营血证候与三焦证候作为证候类下的基本的锚点，进行类与类层次划分。此外，古籍中的词义辨析、合并与归类、一词多义的处理[12]是进行概念抽取、确定领域本体核心概念的关键步骤，但是从整体来看各个学科有着自己的特点。例如对针灸本体进行扩展时，只需要在层次结构的某些分支下增加新的概念；但是针对中医诊断中证候、症状的古语描述，比较难处理；中医古籍中一些特殊的内容，例如“祝由”，概念的抽取无规律可循。

4 结论

4.1 基于本体构建中医古籍知识库切实可行

本研究初步探讨了基于本体建立中医古籍知识库的方法、工具、实现步骤等。文献研究证明本体论通过对领域知识的概念化说明，采用框架系统对概念及其关系进行描述，是一种可行的中医古籍知识表达方法。选取 7 步法，应用 Protégé 3.0，既能够表示知识的等级结构，也可以表示知识的组成结构。

4.2 需探索自动或半自动构建方法

中医本体的构建是一个复杂的过程，纯手工的中医本体构建不但需要领域专家的参与，而且工作量巨大，极易导致知识获取的瓶颈。并且本体中获取的知识需要不断地更新，单靠手动的方式构造，容易造成信息的过时。因此找到一种自动或半自动的方法，利用现有的本体和从已有数据库、网络资源中提取的本体完成中医领域本体的构建是一件极具现实意义的工作。

4.3 基于本体构建中医古籍知识库的重要意义

应用本体论作为中医古籍知识表达的理论和方法，构建中医古籍领域本体，有利于突破古籍深度利用的瓶颈，有利于改善和解决中医古籍数字化进程中所面临的问题，为中医古籍专家系统、信息检索系统等研究提供智能基础，最终达到推进中医药信息学的发展，实现中医古籍知识的采集、存储与利用，促进领域知识共享的目的。

参考文献

1 Gruber T R. A Translation Approach to Portable Ontology Specification〔J〕. Knowledge Acquisition, 1994, 5 (2): 199 –220.

2 冯广义. 中医古籍的整理研究与中医学发展〔J〕. 中医药导报, 2007, 9 (13): 10 –14.

3 张雪梅. 古籍数字化与文献信息资源共享〔J〕. 天津工业大学学报, 2002, 3 (21): 85 –86.

4 Neches R. Enabling Technology for Knowledge Sharing〔J〕. A I Magazine, 1991, 12 (3): 36 –56.

5 Studer R. Knowledge Engineering: principles and methods〔J〕. IEEE Transactions on Data and Knowledge Engineering, 1987, 25 (112): l61 –197.

6 杜文华. 本体的构建及其在数字图书馆中的应用研究〔D〕. 武汉: 武汉大学, 2005: 20 –22.

7 章学诚. 校禅通义·校摊条理, 章学诚遗书〔M〕. 北京: 文物出版社, 1985: 98.

8 王德禄. 知识管理的 IT 实现〔M〕. 北京: 电子工业出版社, 2003: 86 –90.

9 程佳羽, 史睿. 古籍数字资源的知识库建设解析〔J〕. 数字图书馆论坛, 2006, 12 (31): 1 –3.

10 高成勉, 包含飞, 周强. 本体构建原则及其在中医顶层本体构建中的应用〔J〕医学信息, 2008, 21 (5):: 581 –583.

11 李新霞. 基于本体的中医学脾胃病知识库的构建〔D〕. 南京: 南京理工大学, 2008: 33 –34.

12 孙海舒, 李斌, 王蕊, 等. 中医古籍书目数据库标注中若干问题的探讨〔J〕. 中国中医药信息杂志, 2007, 10 (14): 103 –104.