

医学信息学本体知识库系统手工构建研究

宣腾^① 李科* 曾东^① 周焕来^①

摘 要 自2010年我国推行医改政策以来,医疗信息化行业发展迅速。传统的信息技术已不能充分满足医疗信息行业的发展需要。此时CDA技术、本体与语义技术逐渐引入医疗信息化行业的应用中来,涉及临床信息采集、处理、管理和使用等多方面。以由卫生部颁布的《卫生信息数据元目录》为基础,设计开发出基于标准数据元的医学信息学本体知识库,并构建以CDA Schema技术为基础的数据采集引擎,致力于解决当前区域医疗中临床信息合理化采集和整合使用的问题。

关键词 本体构建技术 CDA Schema技术 临床信息管理 语义应用

Doi:10.3969/j.issn.1673-7571.2013.08.010

Research of Medical Informatics Ontology Knowledge Base Manual Construction / XUAN Teng, LI Ke, ZENG Dong, et al//China Digital Medicine.-2013 8(8): 33 to 36

Abstract The health care industry developed rapidly since Chinese government has introduced the new deal in this area. Traditional technology cannot meet the needs of the development of this area. New technologies like CDA Ontology and Semantic web have been used in kinds of applications, such as data collection data management and later applications. This paper design and build a new system to solve the problem that we cannot easily collect and deal with the clinical data produced in daily medical treatment. This system include an ontology base which base on <Health information data directory>, and a CDA data mapping engine.

Keywords ontology construction technology, CDA Schema technology, clinical information management, semantic application

Fund project The National Science & Technology Pillar Program (No. 2012BAH07F01); Key Projects in Sichuan Province Science & Technology Pillar Program (No. 12ZC0220); The Fundamental Research Funds for the Central Universities (No. ZYGX2011J096)


Corresponding author Associate Professor, School of Life Science and Technology, UESTC, Chengdu 610054, Sichuan Province, P.R.C.

1 引言

根据相关机构调研结果显示,我国医疗卫生信息技术(HCIT)系统几乎都是为现有医疗服务系统按专业划分的组织结构,导致绝大部分HCIT系统变成无法互联互通的“信息孤岛”,病人信息被分散在各系统中,不能由外部进行访问,也很难被整合在一起,使医生大部分时候是在没能掌握病人完整信息的情况下进行诊断和治疗。这也是大部分重复检验检查、不完整治疗和医疗事故的根源。

我国卫生部颁布的《卫生信息数据元目录》为各医疗信息系统提供了统一的数据元标准,是对区域医疗机构中产生的临床数据进行整合的前提和基础。基于《卫生信息数据元目录》开发的医学信息学本体知识库系统将为区域医疗信息的整合带来更为安全有效的医疗服务,覆盖整个医疗周期各个活动的信息以及安全可靠的病人数据检索查询功能将大幅提高病人就诊效率和质量。

基于以上问题,主要讨论本体工程技术在医疗信息领域中的应用。即如何使用国家标准数据元来构建区域内通用的

 **基金项目:** 国家科技支撑计划项目(编号:2012BAH07F01);四川省科技支撑计划重大项目(编号:12ZC0220);中央高校基本业务费项目(编号:ZYGX2011J096)

*通讯作者:电子科技大学生命科学与技术学院博士,副教授,610054,四川省成都市成华区二段四号

①电子科技大学生命科学与技术学院,610054,四川省成都市成华区二段四号

医疗信息本体知识库系统；使用CDA Schema技术实现区域内临床数据的采集、上传和管理功能；通过对语义技术的支持，实现对医疗信息数据更深层次的分析与利用。

2 基于医学信息学本体知识库的数据处理系统

不同于以往本体知识库的开发模式，采用基于《卫生信息数据元目录》的国家标准化数据元进行本体知识库的手工开发模式。以CDA Schema技术开发构建本体知识库的数据接口，实现区域内临床数据的采集、上传、存储和语义应用等数据处理功能。

文中提出的医学信息学本体知识库系统由医学信息学本体知识库、前端CDA Schema临床数据映射引擎和WEB SERVICE语义应用服务三部分构成。集合了区域内对临床数据处理的各项功能，并提供一定程度上的可复用性和后续开发接口。系统总体设计架构见图1。

以医学信息学本体知识库为系统核心组件，由CDA Schema映射引擎和WEB SERVICE作为数据接口和服务接口，提供临床数据采集、上传、管

理等服务和区域网络内的临床数据语义检索功能。下面对医学信息学本体知识库进行详细介绍。

医学信息学本体知识库是在实现区域医疗信息化互联互通基础上的进一步信息化、本体化的语义层级应用，在设计和构建时遵循以下原则：复用现行国家数据元标准，满足区域内信息互操作要求，并在此基础开发连接底层关系型数据库的数据接口；通过JENA开发语义检索接口，并发布系统顶层web service服务；构建本体知识库概念元层级结构时应遵循基本的医学、医理逻辑。医学信息学本体知识库功能结构见图2。

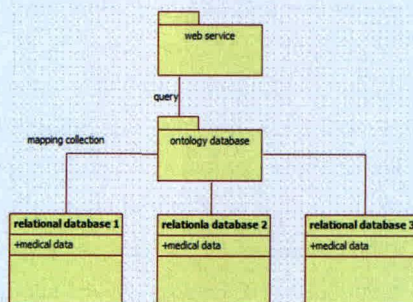


图2 医学信息学本体知识库功能结构

作为医学信息本体库知识库的核心思想，基于《卫生信息数据元目录》进行开发，在首先确保区域内语义互操作性的前提下，进行后期的语义功能开发和数据处理。

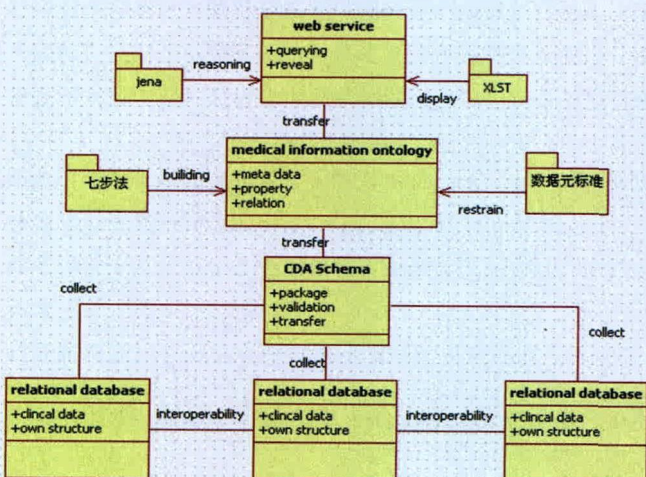


图1 医学信息学本体知识库系统构架

元根据斯坦福大学的七步法来构建本体概念，需经过定义概念、设定概念唯一标识符、限定概念值域范围、创建数据元属性与构建数据元之间医学逻辑关系等步骤。

整理和构建本体知识库概念内容时，采用《卫生信息数据元目录》中的层级结构对概念进行一级分类，构建概念与概念集之间的继承关系，形成推理逻辑树的一级分类（见图3）。



图3 本体知识库层级结构

采用国家标准数据元目录复用形式来构建本体知识库的主要工作在于对数据元进行整理和分类，同时对数据元所含内容进行有效精简，保证本体知识库结构的健壮性和高度结构化。

3 CDA Schema映射引擎与JENA语义检索服务

3.1 CDA Schema数据映射引擎 采用CDA Schema技术开发的底层关系型数据库数据映射引擎在本体知识库系统中，起到采集、上传数据、实例化本体概念的关键作用。其具体结构见图4。

如图4所示CDA Schema映射引擎采用模块化组件，各模块间通过include和import两个功能标签进行关联，组成单一的XML Schema对卫生信息数据元数据集进行整体约束，易于扩展和维护。其中各部分文件分别为：CDA Schema的基本数据集.XSD；CDA Schema自定义数

知识库中使用的概念来自对《卫生信息数据元目录》复用，通过多次筛选和删减，节选出精简的知识库概念。概念分类由《卫生信息数据元目录》中的十七个大类进行压缩和合并，整合为十三类。整理出的数据

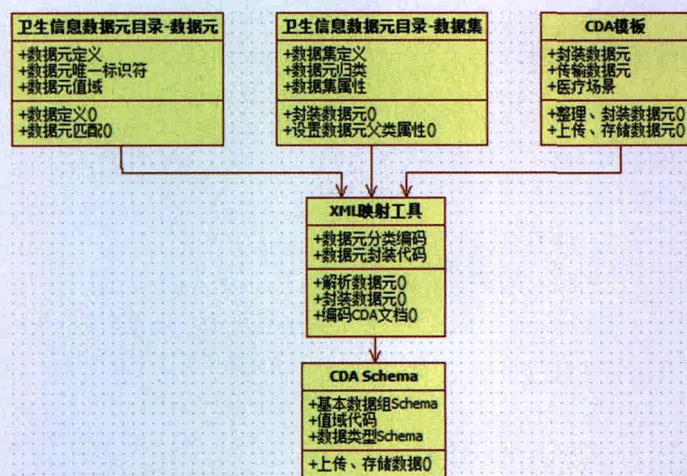


图4 CDA Schema映射流程

据类型.XSD; CDA Schema值域代码.XSD。

数据元定义: CDA Schema中将直接赋值的数据元组件定义为简单类型元素,使用<xs:element name=""><xs:SimpleType>或者<xs:element name=""/>进行声明,对包含子组件的组件,如数据组定义为复杂类型元素,用<xs:element name=""><xs:complexType>声明。

数据类型定义: 使用XML Schema中规定的数据类型加上样式与长度约束来进行定义,构成卫生信息数据元目录中的数据类型和CDA Schema中依照数据类型进行表示的自定义扩展数据类型。

数据属性及数据元注释: 在XML中使用annotation对数据元属性和重要注释进行描述,如“</xs:annotation>...</xs:annotation>”。

数据值域定义: 在XML中使用<xs:enumeration value="" />来声明。

CDA Schema组件处于本体知识库与底层关系型数据库之间,针对不同的数据库在数据采集时,需要使用配置向导。临床信息数据采集时的配置部分主要包括:医疗机构信息、医

疗信息系统数据库信息连接信息以及数据库表结构信息。

3.2 基于JENA技术的语义检索服务 在整个医学信息学本体知识库系统架构中,语义查询系统处于顶层的应用层面,是对本体知识库实际应用功能的开发和展示。主要用于对区域内的临床医疗数据进行语义检索,提高检索准确度和检索效率。文中采用现有JENA API作为语义查询工具。图5为语义查询流程。

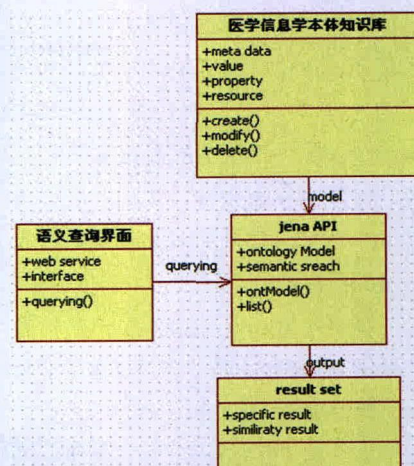


图5 语义检索服务组件结构

在导入本体知识库生成jena模型后,查询功能主要依靠在Model和Resource接口中执行list()来实现。对满足查询条件的类、实例和属性均可通过list()方法获得。文中使用的查询

方法是Model.listStatements(Resource s,Property p,RDFNode o),所有参数的缺省值都为null,作为通配符与任意数据相匹配。

通过jena API能够快速导入本体库进行语义查询。除能查询到符合特定条件的类以外,还可检索出语义相近或者属性相同的类、实例。相比传统的检索模式,语义检索的准确度更高,返回结果集的关联性更强。

4 本体知识库系统对比分析

在医疗卫生领域,缺乏本体技术、语义技术等前沿计算机网络技术应用。但在为数不多的医疗信息领域本体知识库中,不乏兼具创新性和实用性为一体的优秀典范。以下从几方面面对由《生物医学知识整合论》作者包含飞教授主持开发的3Hontology(3H指三高,即血脂、血糖和血压)本体知识库与本文所构建的医学信息学本体知识库系统进行对比分析。

4.1 数据元对比 本文构建的医学信息学本体知识库基础国家数据元标准《卫生信息数据元标准》开发,通过对国标数据元的筛选与复用生成本体知识库中的概念(meta data),而3Hontology本体知识库概念由自己的医学团队开发。采用不同的数据元导致两个本体知识库具有截然不同的性质和功能(见表1)。

表1 数据元对比分析

对比项	医学信息学本体知识库	3Hontology 本体知识库
数据元涵盖范围	依照国家标准、齐全	针对 3H 病种、范围较小
数据元颗粒度	用于数据传输、颗粒度大	颗粒度小
数据元规范性	符合国家标准、高	由医学专家提出、一般
数据元通用能力	符合国家标准、强	针对 3H 病种、无
数据元冗余度	经过反复筛选、低	由专人筛选、一般

4.2 数据元属性对比 医学信息学本体知识库因需要在区域内进行有效的数据互联互通,所以为每项数据设置了完整的定义、标识符和值域范围等,

在数据交互时起到关键作用的属性项。3Hontology本体知识库则由于将工作和研究重点放在数据间的逻辑推理关系上,较为忽略这部分内容(见表2)。

表2 数据元属性对比分析

对比项	医学信息学本体知识库	3Hontology 本体知识库
数据元属性完整性	依照国标、完整	缺乏
是否具有唯一标识符	有	否
是否具有明确定义	有	否
是否标明数据元值域	是	否
数据元属性是否规范	依照国标、规范	否
数据元属性通用性	强	无

4.3 本体知识库逻辑结构对比分析 本体知识库的逻辑层级结构决定了该本体库的语义推理能力,以及其后续的科研潜力。医学信息学本体知识库同3Hontology本体知识库由于解决的问题及服务对象不同,在这方面也有巨大差异(见表3)。

表3 逻辑结构对比分析

对比项	医学信息学本体知识库	3Hontology 本体知识库
数据元分类数目	13	5
数据元分属数目	2	6
数据元逻辑关系复杂程度	较简单	复杂
本体知识库语义推理能力	一般	强

4.4 应用与后续开发能力对比 面对不同的服务对象、采用不同的数据元、针对不同的应用领域,医学信息学本体知识库与3Hontology本体知识库间应用能力的对比没有高低之分。医学信息学本体知识库将应用重点放在如何实现区域内底层临床数据的语义检索上,而3Hontology本体知识库则更重视三高病症的病理逻辑推导和慢性病特征研究(见表4)。

表4 应用与后续开发能力对比

对比项	医学信息学本体知识库	3Hontology
数据兼容性	兼容区域内的数据元,强	基本不兼容,弱
本体知识库通用性	在区域内广泛适用	针对专业医学科研工作者
本体知识库逻辑推理能力	支持基本的语义关联、弱	支持医学逻辑推理,较强
本体知识库的复用能力	符合国标、复用性好	内部实验室开发,数据元缺乏标识符和定义,复用性差
本体知识库后续应用	基于国标后续开发和应用较为容易	由于数据元缺乏标准支持,同时冗余度较高,后续开发能力较弱
生成实例情况	通过导入临床数据自动生成	需要手工录入,容易出错

4.5 对比小结 3Hontology本体知识库包含三高病领域内几乎所有涉及到的数据元和数据元关系类型,具有医学

信息学本体知识库无法比拟的专业性和逻辑推理能力。通过对比数据元和数据元分类可以看出,3Hontology本体知识库并不具备兼容区域内关系型数据库的能力,无法通过自动导入临床数据生成实例,而需通过手工录入数据。这会造成大量人工成本,也容易出现差错,同时3Hontology本体知识库涉及领域范围狭窄,适用于对特定病种进行专门性研究使用。医学信息学本体知识库因采用国家通用的数据元标准,具备在实行《卫生信息数据元目录》标准的区域内数据互联互通、语义互操作能力,在专业性和推理能力上有所欠缺,但涵盖数据面广,后期应用开发潜力大。

5 总结及展望

研究了医学信息本体知识库的构建问题,创新性的以《卫生信息数据元目录》、《卫生信息数据元值域代码》为基础,深入学习和研究了国际上同类型的医疗信息数据交换标准与领域本体知识库的构建方法。开发了基于《卫生信息数据元目录》的医学信息学本体知识库,并采用CDA Schema技术开发底层数据采集和封装、传输的映射引擎,实现对区域内接入中心信息平台的各关系型数据库的语义检索功能。

针对区域内医学本体知识库构建问题进行了研究,接下来重点是增加医学、医理逻辑推理属性,满足专业用户在语义推理方面的需求,同时考虑更多本体知识库的语义功能,如医嘱自然语言解析和临床决策支持等。

参考文献

- [1] 陈云志,刑美国,陈琦,等.基于本体的疾病知识库设计[J].中国数字医学,2010,5(10):29-31.

[2] 于彤,赵阳,崔蒙,等.语义网技术在生物医学中的应用现状及发展趋势[J].中国数字医学,2012,7(10):9-12.

[3] Ifikhar S, Nawaz F, Ahmad HF, et al. Efficient discovery of OWL-S based HL7 compliant healthcare web services profiles. 6th International Conference on Emerging Technologies. IEEE Computer Society, 2010: 287-292.

[4] 李鹏飞,黄冉,姚琴,等.面向医学信息交换的语义查询系统设计[J].中国数字医学,2012,7(12):24-27.

【收稿日期:2013-05-07】

(责任编辑:赵士洁)

业界观察

青海:基层卫生机构信息化建设启动

近日,青海省卫生计生委启动基层卫生机构信息化建设项目,旨在进一步加快基层信息化建设步伐,确保医改各项任务顺利完成。

该省基层卫生机构信息化建设项目,是国家卫生信息化“3521”工程的一部分。2012年以来,国家持续加大卫生信息化建设投入力度,累计为该省投入资金8500万元,推进基层卫生机构信息化项目建设。

该省制订了基层卫生机构管理信息系统建设实施方案,计划建设16个二级数据中心,并为基层卫生机构配备核心软硬件设备;完成各乡镇卫生院、社区卫生服务数据中心(站)网络线路及局域网改造建设工作,开展各县区数据交换平台建设。计划到今年年底,基层卫生机构业务信息系统上线运行率达到90%,实现对基层卫生机构业务运行的动态监管与基本公共卫生、基本医疗数据的互联互通。

(来源:健康报)

OBSERVATION