

# 基于XML的电子病历结构化存储研究

陈 可

(南京市胸科医院 信息科, 江苏 南京 210029)

**摘 要:**针对电子病历系统中病人数据如何格式化处理、信息如何交换与共享等问题,展开了基于XML的电子病历数据存储研究。该研究以病人为根的树状结构化的数据结构为分析对象,并借助DTD(文档类型定义)标准定义该数据结构,克服了为适应病历中不同内容结构的变化,而改动程序的传统做法。该文选取DB2作为电子病历的数据管理平台,利用其兼具处理关系数据和XML数据的特点,实现了病历信息的结构化存储及交互,为分布异构系统之间临床信息集成奠定了基础。

**关键词:**电子病历;可扩展标记语言;文档类型定义;结构化存储

中图分类号:TP315

文献标识码:A

文章编号:1673-629X(2012)05-0205-04

## Research on Structure Storage of Electronic Medical Records Based on XML

CHEN Ke

(Information Department, Nanjing Chest Hospital, Nanjing 210029, China)

**Abstract:** According to the problems as how to deal with patients data formatting, and how to transfer and share information in electronic medical record system, the data modeling method based on XML was raised by this paper. By means of data structure defined by DTD, this method analyzed the data structure rooted by the patient, which to overcome modifying code when someplace needed to change in electronic medical record system. DB2 was selected to be data management platform for the electronic medical record system. By adopting the characteristics of DB2 with relationship data and XML data processing, the medical records are structured stored in DB2, which also lays a foundation of clinical information integration in distributed and heterogeneous system.

**Key words:** electronic medical records; extensible markup language; document type definition; structure storage

### 0 引 言

随着医疗信息化的深入发展,医院信息系统已由传统的医院管理信息系统(Hospital Management Information System, HIS)向临床信息系统(Clinical Information System, CIS)转变,电子病历(Electronic Medical Record, EMR)更是其中的基础和重点<sup>[1]</sup>,它的建立是以病人诊疗信息为主线,集成了病人诊疗相关的各个环节,涵盖了医技检查、检验、医嘱与护理记录等信息。随着病历信息量的辐射面越来越广泛,病历的结构化信息会随之增多,而病历的结构化程度也会变得更为细化,这就要求EMR系统在结构上必须要保持病历的整体性<sup>[2]</sup>。因此,在设计EMR数据存储结构时,须将这一发展要求纳入考虑,使系统能够随时将新出现的结构化内容集成到病历中来。

可扩展标记语言(Extensible Markup Language,

XML)为这一发展创造了条件。同时也迎合了HIS采取的医院信息系统电子数据交换标准(Health Level Seven, HL7)<sup>[3]</sup>。以检查报告为例,心脏超声与腹部超声的报告在格式上会存在一定的差别,对于这一差别的处理,传统的数据库模式是将在医院业务流中产生的各种检查报告抽象成一种统一的结构,但这却无法体现彼此间的区别,要解决这一问题只能依靠修改应用程序来实现。而在XML方式下,可将它们设计成不同的文档类型定义(Document Type Definition, DTD)<sup>[4]</sup>,再将新设计的报告集成到系统中,无须修改已有的软件。显然,相对于传统数据库方式,这种依托XML技术的数据存储模式更适合于描述病历这种复杂的内容,也更游刃有余地处理病历中出现的结构或内容的变化。

### 1 基于XML方式的电子病历结构化存储

#### 1.1 电子病历系统的数据逻辑结构

EMR研究的基本问题是病历的描述模型,这可以

借助 XML 技术来描述。作为一种标识语言,更是一种可以定义描述对象结构的元语言,XML 可以将网络上传输的文档规范化,并赋予标记一定的含义,使系统间交换的信息可以互相“理解”<sup>[5]</sup>。

通过分析病历的业务流,我们发现,EMR 的数据具有树型结构的特点,它所强调的是以病人为中心的管理信息系统,建立以病人为中心的数据归档。当我们用 XML 文档描述病历内容时,“病人”就相当于 XML 文档的根,而病人的所有相关信息则视为根的分支,整个描述过程完全是以“病人”为中心展开的,如图 1 所示。

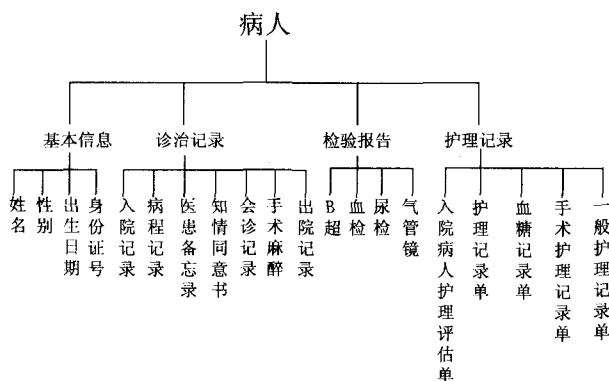


图 1 EMR 系统的树型数据结构

## 1.2 基于 XML 电子病历数据模型构建方法

临床文档是 EMR 中各类业务活动记录的基本形式。如:病程记录、表格式专科病历等,这些病历的结构内容各不相同,但其中有些数据则存在着一定的层级结构,有包含与被包含的关系,也有按同类属性相互嵌套的关系。比如,首次病程中有很多信息来源于入院记录的主诉、现病史及辅助检查;再比如:医生在撰写病历过程中,自行建立的各种结构化病历模板,其中会存在信息相互嵌套的关系。这些病历文档若采用以文本为单位的存储模式,不仅难以实现病历之间的索引与比较,还会容易出现数据不一致的现象。因此,选择基于 XML 的方式建立数据模型,将一份病历划分为一个单据,而每份单据由若干标题组成,每个标题则由若干节段组成,对于每个节段可由若干关键词或各单据可能共享的信息组成,当然,每个节段又可包含若干标题,然后标题再包含节段,如此循环嵌套。

目前,基于 XML 技术建立的数据建模手段有两类,文档类型定义 DTD 及 XML 大纲<sup>[6]</sup>,论文采用前者数据建模手段,通过一张自引用表来实现自引用,基于建立的 DTD 模型,将各份病历单据中的关键词及各单据可能共享的信息全部设计为一个一个的节点,通过 XML 语言中 XSL 的格式化描述功能对各个节点设计一种表现方式<sup>[7]</sup>。这样一来既保持了病历文档的属性结构,又通过 XML 语言的自我描述实现了整份病历的

结构化存储,体现了数据描述从以“类型”为中心到以“病人”为中心的转换。而且,因为每次操作的是其中一份单据的某个节段,所以,一份病历在录入、编辑和保存时的速度不会因此而受影响。

考虑到结构化数据存在的异构性,为了使数据具有较强的灵活性及适应性,需对 DTD 文档中的标记符号做统一定义,DTD 标记符号的定义如下所示<sup>[8]</sup>。文中对每个子结构所允许出现的数目、出现的顺序等也都做了标注。

加号(+):说明该标记至少要出现一次;

问号(?):说明该标记可以出现零次或一次;

星号(\*):说明该标记可以出现一次或多次,或者不出现;

逗号(,):用于在顺序列表中分隔元素;

管道符(|):用于可选的子标记;

无符号标注:表明必须出现且只可出现一次。

在编写 DTD 时,如果某一元素不含子元素,则要在其后以关键词#PCDATA 标注,如若某一元素的属性值必须出现,可用#REQUIRED 表示。通过分析 EMR 数据结构(如图 1 所示),得知病历号是病人必须含有的属性,必须显示,是以一串文本数据来描述的,故属性类型为 CDATA。因此,编写的 DTD 文档如下:

```
<? xml version="1.0" encoding="GB2312"? >
```

```
<! ELEMENT 病人( 基本信息,诊治记录+, 检验报告?, 护理记录)>
```

```
<! ATTLIST 病人 病历号 CDATA #REQUIRED>
```

```
<! ELEMENT 基本信息(姓名,性别,出生年月)>
```

```
<! ELEMENT 诊治记录(入院记录,病程记录,医嘱备忘录+,知情同意书*,会诊记录*,手术麻醉*,出院记录)>
```

```
<! ELEMENT 护理记录(入院病人护理评估单,护理记录单,血糖记录单,手术护理记录单*,一般护理记录单)>
```

```
<! ELEMENT 检验报告(B超*|血检*|尿检*|气管镜*|)>
```

对于系统中一些重要的单据,如入院记录、出、入院小结、检验申请及一些专科检查均在关系数据库中建立对应的数据表。经过采集转换之后,形成以病人为中心的 XML 文档,通过 XML 解析模块,并依据 MAP 文件的规则进行映射,将 XML 文档的结构和内容映射到关系数据库中<sup>[9,10]</sup>,并保存数据结果集。这便于支持日常的业务处理和病人信息采集,同时也满足了临床日常所需的数据检索和更新要求。因此,在 EMR 数据存储管理中,采用 XML 文档记录病历,并不排斥病人信息的数据库表示,两者有其各自的适用范围,它们将共同存在于 EMR 数据存储策略中<sup>[11]</sup>。

图2是XML文档的存储原理图。

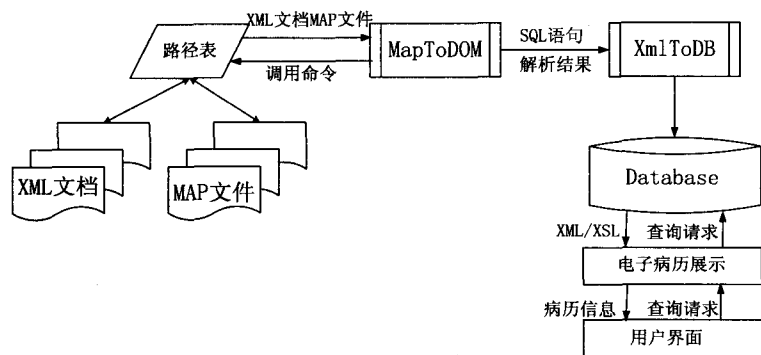


图2 XML文档存储原理图

## 2 关键技术及其实现

### 2.1 病历的XML文档描述

XML文档的编写应遵循DTD规定的格式,并按照下列原则进行:

(1) 标记所需要的XML声明,给出文档需要声明的信息。如:XML规范的版本号、字符编码方式等信息。

(2) 必须存在一个单一的、用于嵌套所有其他元素和内容的根元素。

(3) XML文件中的所有元素、属性和实体必须使用正确的XML语法,严谨的结构化语法结构可以减少浏览器的解析错误。

以EMR系统中医生自行定义的病历模板为示例,依据上述XML文档编写原则,给出一段部分XML文档,文档保存为“template.xml”,如下所示:

```
<? xml version="1.0" encoding="GB2312"? >
```

```
<! DOCTYPE SYSTEM "template">
```

```
<root>
```

```
<病史>
```

```
<TemplateName>现病史</TemplateName>
```

<TemplateContent>既往体健,否认“肝炎”、“伤寒”、“结核”等传染病史,否认高血压、糖尿病史,无药物及食物过敏史,无重大手术外伤史,否认输血史,预防接种史随社会。

```
</TemplateContent>
```

```
<TemplateName>个人史</TemplateName>
```

```
<TemplateContent>……. </TemplateContent>
```

```
<TemplateName>家族史</TemplateName>
```

```
<TemplateContent>……. </TemplateContent>
```

```
</病史>
```

```
</root>
```

### 2.2 MAP文件的编制

对于XML文档与数据库关系表在结构上存在的

差异问题,可通过编写MAP文件来解决。在MAP文件里,定义了这两种数据结构之间的映射规则,进而可实现数据从XML文档到数据库表的转移。

```
<ClassMap>
```

```
<ElementType name="病史"/>
```

```
<ToClassTable name="medical tem-  
plate">
```

```
……
```

```
</PropertyMap>
```

```
<ElementType name="个人史">
```

```
<ToTable name="medical template">
```

```
<ToColumn name="个人史">
```

```
</Property>
```

```
</PropertyMap>
```

```
<ElementType name="家族史">
```

```
<ToTable name="medical template">
```

```
<ToColumn name="家族史">
```

```
</Property>
```

```
</PropertyMap>
```

```
……
```

```
</PropertyMap>
```

```
</ClassMap>
```

图2所示的XML文档存储原理图反映了从XML文档到关系数据的映射规则,图中的MapToDOM子模块通过连接数据库,读取XML文档和MAP文件的存放地址及其各自相应的文件,借助解析器对XML文档进行解析,并建立XML文档的对象模型树(DOM)。XML文档和MAP文件经解析后的数据提交给XMLToDB模块,按照既定的映射规则实现XML文档到数据库表的映射。

## 3 EMR系统的数据管理策略

病历的XML描述与关系数据库记录各有其适用范围,它们在EMR的数据管理中并存。作为EMR数据管理平台,文中选用IBM公司开发的DB2 9.5。这是一个混合数据库管理系统,其编译器可同时操作关系数据和XML数据。关系数据主要是指医院日常的业务处理及病人信息的采集,涉及到大量的数据检索与更新;XML数据体现在EMR系统所实现的病历浏览,为了保证业务系统的高效率运行,过期数据不宜在数据库中长期保存,因此,在病人出院后,病历内容会被转化为XML描述文件。在DB2中,XML数据类型的描述是将节点以树的形式展现出来,解析后的整个XML文档存储于关系数据库中;当然,也可在保存XML文档层次结构的前提下,在节点粒度上存储XML

文档的片段,而不破坏其层次结构。当其中一部分数据发生变化时,不需要重发整个结构化的数据<sup>[12]</sup>。这不仅避免了重建 WEB 页,还减轻了服务器的负担,这种数据更新方式称之为“粒状更新”。基于 DB2 9.5 兼具有处理 SQL 和 XQuery 的两种不同解析方式的特点,便于对关系数据和 XML 数据的灵活处理,故选其为 EMR 系统的数据库管理平台,DB2 的数据存储图如图 3 所示。

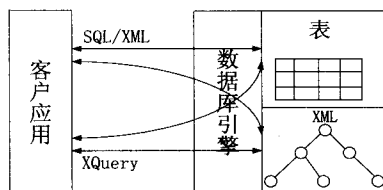


图 3 SQL 或 XQuery 透明查询

#### 4 结束语

论文将 XML 技术应用于医疗临床系统,定义了以病人为中心的 EMR 数据结构,重点论述了 EMR 的数据存储策略。由于 XML 关系数据库的模式是二维的表结构,因此相应的存储策略就是将 XML 文档的树形结构转换为表结构。

文中以 DTD 来定义需要交换的 EMR 文档,通过验证文档的有效性,保证文档内容遵守共同的定义和规范,一方面,实现了病历数据的存储和展现,另一方面,DTD 的集成能将病历中新出现的结构化内容集成到 EMR 系统中。实践应用表明,XML 技术实现了病历的自由输入和结构化输入相结合的可行性,这一数据模型,可作为 EMR 与医院内部其他业务数据库数据

交互的标准,为实现整个医疗信息平台的数据集成奠定了基础。

#### 参考文献:

- [1] 赵洋,李万龙,白杰英.基于本体的电子病历检索系统研究[J].计算机技术与发展,2010,20(3):211-217.
- [2] 覃永胜,林崇健.基于 DB2 pure XML 的电子病历研究[J].中国数字医学,2009(2):14-17.
- [3] Health Level Seven (HL7) Version 3.0[S]. Michigan: Health Level Seven Inc, 2003.
- [4] 周健,孙丽艳.面向对象 XML 的存储模式的研究[J].计算机技术与发展,2009,19(3):114-119.
- [5] Graves M. Designing XML Database[M]. 北京:机械出版社,2002:46-47.
- [6] Erdman M, Studer R. How to Structure and Access XML Documents with Ontologies[J]. Data and Knowledge Engineering, 2000,36(3):317-335.
- [7] 庄子明.基于 XML 的数据库技术及应用[J].计算机工程,2002,28(6):119-121.
- [8] 李德有.基于 XML 电子病历数据存贮与转换的研究[D].哈尔滨:哈尔滨理工大学,2007.
- [9] 方翔.XML 文档到关系数据库的直接转换[J].计算机工程,2001,27(1):65-66.
- [10] 曹亮,王茜,卢著.XML 数据在关系数据库中存储和检索的研究和实现[J].东南大学学报(自然科学版),2002,32(1):124-127.
- [11] 张继光.电子病历特征及其设计方案探讨[J].中国医院管理,2003,23(5):44-46.
- [12] Widom J. Data Management for XML: Research Directions[J]. IEEE Data Engineering Bulletin, 1999,22(3):44-52.

(上接第 204 页)

#### 参考文献:

- [1] 李玉山.数字视觉视频技术[M].西安:西安电子科技大学出版社,2005.
- [2] 严云洋,郭志波,杨静宇.人脸识别特征抽取方法的研究进展[J].淮阴工学院学报,2007,16(3):24-30.
- [3] 刘党辉,沈兰荪.视频运动对象分割技术的研究[J].电路与系统学报,2002,7(3):77-85.
- [4] Nilsson N, Nordberg J, Claesson I. Face Detection Using Local SMQT Features and Split Up Snow Classifier[C]// ICASSP. [s.l.]:[s.n.], 2007:589-592.
- [5] Nilsson M, Dahl M, Claesson I. The successive mean quantization transform[C]//IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP). [s.l.]:[s.n.], 2005:429-432.
- [6] 王刚,周激流,何坤.基于 AdaBoost 和 LLE 的视频人脸特征提取研究[J].四川大学学报,2008,45(3):512-516.
- [7] 朱杰,唐振民. ASM 与彩色 Gabor 特征相结合的人脸关键特征点提取[J].计算机科学,2001,37(4):265-268.
- [8] 李晓卫,杨邦荣,杨炼.改进的用于人脸面部特征定位的 ASM 方法[J].电脑与信息技术,2008,16(2):14-17.
- [9] 刘洵,张大力.基于 ASM 的 CT 图像序列标记点定位方法研究[J].计算机工程与应用,2005,41(13):180-182.
- [10] Cootes T F, Taylor C J, Cooper D H, et al. Active shape models: Their training and application[J]. CVGIP: Image Understanding, 1995, 61:38-59.
- [11] 刘党辉,沈兰荪.视频运动对象分割技术的研究[J].电路与系统学报,2002,7(3):77-85.
- [12] 俞扬信,严云洋.视频序列中的人脸检测与定位算法研究[J].计算机技术与发展,2009,19(2):109-111.