

元数据研究在医学数据资源共享中的作用

胡凯¹, 刘丽华^{2a}, 徐勇勇¹, 尹岭^{2b}

[摘要] 目的 在医学数据资源共享中,通过对元数据的研究对异构医学数据资源进行检索与发现。方法 研究元数据的定义与特征,制定元数据标准,然后建立目录服务系统进行数据检索服务。结果 建立了基于元数据的目录服务体系,实现了数据发现服务。结论 元数据方法的研究是对异构数据进行检索与发现的有效方法,也是实现医学数据资源共享的重要途径。

[关键词] 元数据;数据集;数据检索

Metadata Study in Medical Data Resource Share HU Kai, LIU Li-hua, YIN Ling, et al. Department of Health Statistics, the Fourth Military Medical University, Xi'an 710032, Shaanxi, China

[Abstract] Objective To study the application of Metadata in search and find data with different structure in medical data resources share. Methods Study the definition and character of metadata, establish the metadata standard and directory service system to execute the data search service. Results The Directory Service System based on the method of metadata has been developed, in which the data search service can be realized. Conclusion The Directory Service System based on the method of metadata has been developed, in which the data search service can be realized. Conclusions Its a effective way to search and find data with different structure using method of metadata, and it is a important way to realize medical data resource share.

[Key words] metadata; dataset; data search

中图分类号: R197.324 文献标识码: A 文章编号: 1006-9771(2006)05-0454-02

[本文著录格式] 胡凯,刘丽华,徐勇勇,等.元数据研究在医学数据资源共享中的作用[J].中国康复理论与实践,2006,12(5):454—455.

各种信息系统的广泛应用为医学领域带来了庞大的电子数据资源。要对这些存在于不同结构和不同格式数据库中的数据资源进行交流与共享,经常要面对的一个问题就是如何让用户在各种异构数据集群体中进行检索和数据发现。一个有效的途径就是同过元数据的研究来实现数据的发现服务。

1 数据发现服务对元数据研究的需求

1.1 共享数据集 我们把需要进行共享的数据资源的最小粒度定义为数据集(dataset),即可以标识的数据集合。

数据集是一个相对抽象的概念,它与具体的数据库概念不同,只要满足可以进行描述、具有一定意义的数据集合就是数据集,用于数据共享的数据集是数据共享的基本内容和最小粒。

1.2 数据集共享对元数据的需求 由于数据集是共享服务研究的最小粒度,数据的交流与共享活动是不会进入数据集的内部对各个数据集不同的数据内容、数据格式、存储方式以及数据量等属性信息等进行获取和处理的。这就需要一种对数据集的整体进行统一、规范描述的信息来提供给共享服务系统。这样,共享服务系统只要获取到这个整体性的描述信息就可以实现检索与数据发现等共享服务操作了。

这个对数据集的整体进行描述的信息就是元数据。

2 元数据研究

2.1 元数据的定义与特征 元数据(metadata)的定义是“关于数据的数据(Data About Data)”。可以用下面的例子来说明元数据的定义与特征。

例如,一个科研人员完成了他的科研任务,需要将自己的研

究数据上交由本单位的档案管理部门管理。他将所有的数据及研究资料存储入一个光盘上交给资料管理部门,而资料管理部门为了便于档案的存放、管理和今后使用中对该资料的查找,要求该科研人员填写一张制式的、通用的表格来说明这张光盘的情况及内容等信息。那么,我们可以认为这张制式表格所要求的内容就是这张光盘的“元数据”。

从以上的例子我们可以看出元数据的一些特点:①元数据是用来说明实体数据的,它被要求能够描述实体数据的内容和概况信息。②元数据内容可以与具体的数据内容、数据格式等无关。③元数据具有对实体数据进行组织、管理和检索的作用。

在医学领域数据共享服务过程中,元数据描述的对象就是医学领域中进行交流与共享的数据集。

2.2 元数据标准的研究 在上面的例子中,用于数据存档时所填写的制式的、通用的表格规定了建立完整“元数据”所需要的内容,这个表格所要求的填写内容与规定就是一个元数据标准的概念。

元数据标准定义了完整描述一个数据集所需要的数据项集合,各数据项语义定义和著录规则等。它提供了关于数据的标识、内容、分发、数据质量、数据表现、扩展、数据模式、限制和维护等方面的信息。

2.3 目录服务模式的研究 通过元数据的研究与标准制定,用户可以通过目录服务体系来获得由元数据技术实现的数据检索与发现服务,从而实现数据的交流与共享。

由于元数据是按照一定标准对信息资源做规范化描述,所以说它是从信息资源中抽取出的相应特征所组成的一个特征元素集合。通过元数据,人们能够对信息资源的进行详细、深入的了解,包括信息资源的格式、质量、处理方法和获取方法等各方面细节。

目录服务系统是通过元数据技术实现的一种信息服务标

作者单位:1. 第四军医大学卫生统计学教研室,陕西西安市710032;2. 解放军总医院,a. 医疗统计科;b. 神经信息中心,北京市100853。作者简介:胡凯(1976-),男,河南洛阳市人,硕士,主治医师,主要研究方向:医院信息资源分析利用、医学信息标准化。通讯作者:刘丽华。

准模式,它通过元数据将信息以动态分类的形式展现给用户,用户通过目录服务系统可以快速确定自己可能所需数据的元数据信息。

通过元数据信息,用户可以进一步确定自己所需要的具体数据资源,然后根据每个元数据记录与它所描述的数据资源之间的导航来访问数据资源。

3 元数据研究的功能实现

3.1 元数据内容标准的制定

3.1.1 元数据标准的制定 首先要建立元数据标准,来界定对各个数据集进行描述的标准化、规范化通用内容。制定的方法参考如下:①根据共享与交流实际情况:根据所要建立共享服务系统所包含资源的专业、范围、内容,以及共享的服务方式、服务对象等制定元数据标准;②参照元数据国际标准:可参照都柏林核心元数据集(Dublin Core Metadata Element Set, Version 1.1: Reference Description, 2003-06-02)与 ISO 19115 地理信息-元数据(Geographic information Metadata)等元数据标准制定元数据标准的基本内容;③参照医学领域标准:如参照 SNOMED JCD-10、DICOM、LOINC 等医学领域标准进行元数据标准的制定。

3.1.2 元数据标准的表达 元数据的表达方式目前采用较多的是元数据摘要表示、字典描述和 UML 图描述的表达方式。

3.1.2.1 摘要表示 摘要表示使用定义、英文名称、数据类型、

值域、短名、注解、子元素和扩展巴氏范式(以上属性并非全部必选)来描述元数据。

例如:

在线资源 < <数据类型> >

定义:可以与负责人或负责单位联系的在线信息

英文名称: onLine Resource

数据类型:复合元素

短名: Online Res

注解:可选项,最大出现次数为 1

子元素:在线资源 =

链接地址 +
0{ 协议 } 1 +
0{ 应用领域专用标准 } 1 +
0{ 名称 } 1 +
0{ 说明 } 1 +
0{ 功能 } 1

扩展巴氏范式: cntOnline Res = linkage, 0{ protocol } 1, 0{ appProfile } 1, 0{ orName } 1, 0{ orDesc } 1, 0{ orFuncnt } 1。

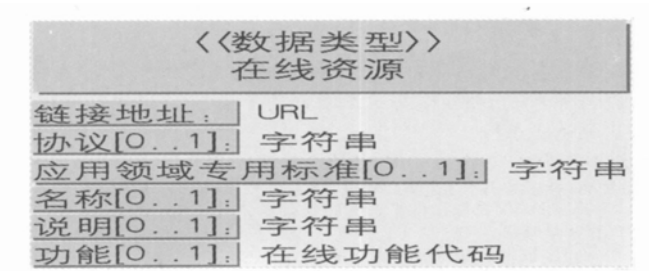
3.1.2.2 字典描述 数据字典以表格的形式描述元数据的特征属性。字典中第 0 行定义元数据实体。数据字典通过名称/角色名称、短名和域代码、定义、约束/条件、最大出现次数、数据类型、域 7 个属性来定义元数据实体和元数据元素。如表 1。

表 1 元数据的字典描述

名称/角色名称 (中文)	名称/角色名称 (英文)	短名	定义	约束/条件	最大出现次数	数据类型	域
0 在线资源	OnLine Resource	Online-Res	可以获得数据集、规范、领域专用标准名称和扩展的元数据元素的在线资源信息	使用参照对象的约束/条件	使用参照对象的最大出现次数	类 < <数据类型> >	第 142 - 147 行
1 链接地址	linkage	linkage	使用 URL 地址与类似地址模式,如 http://www.statkart.no/isotc211/,进行在线访问的地址	M	1	复合型	URL (IETF RFC 1738 IETF RFC 2056)
2 协议	protocol	protocol	使用的连接协议	O	1	字符串	自由文本
3 应用领域专用标准	applicationProfile	appProfile	可以与在线资源一起使用的专用领域专用标准名	O	1	字符串	自由文本
4 名称	name	orName	在线资源名称	O	1	字符串	自由文本
5 说明	description	orDesc	在线资源是什么/做什么的详细文字说明	O	1	字符串	自由文本
6 功能	function	orFuncnt	在线资源功能代码	O	1	字符串	在线功能代码表 < <代码表> > (B 3. 3)

3.1.2.3 UML 图描述 用 UML 中“包”的概念表示元数据子集,用 UML 中“类”的概念表示元数据实体,用 UML“类的属性”的概念表示元数据元素。

例如:在线资源(1属元数据信息)UML 图表示



后,可以通过目录服务系统来实现用户的检索与数据发现服务。

目录系统一般由用户界面、元数据采集、元数据网关、元数据发布服务器等功能模块组成。

3.2.1 用户界面 提供标准接口的人机交互界面,用户可以通过该模块进行各种数据信息查询检索工作。

3.2.2 元数据采集 用于采集和加载各类元数据。用户通过该部分进行数据集元数据内容的著录,并由系统进行元数据在共享服务系统中的注册、发布。要求该部分模块具有多标准支持、辅助编辑、规范性检查等功能。

3.2.3 元数据网管 主要作用是管理各元数据发布服务器,向元数据发布服务器分发用户提交的查询请求和返回查询结果。

3.2.4 元数据发布服务器 用于提供网络信息搜索和提取服务。

3.2 数据目录服务系统的功能实现 元数据的标准内容确定之

(收稿日期 :2006-01-23)