元数据及其在医疗学科数据共享中的应用

河北省儿童医院　　周莲茹
河北省卫生厅　　王坤　程 颖

[摘要]从元数据的来源、定义、结构、类型、特点和作用、互操作、元数据标准以及元数据在医学科学数据共享中的应用等几个方面对目前元数据研究进行综述。
[关键词] 元数据；数据共享；医学综述

1、元数据的基本概况

元数据词最早出现于美国航空航天局的《目录交换格式》手册中【1】，被译为“元数据”或“诠释数据”。其英文定义可理解为“描述数据的数据”或“关于资源的结构化数据”。对于元数据的具体含义虽有不同解释，但一般认为元数据是用于提供某种资源的有关信息的结构数据(如题名、外在表征、位置等)。元数据最早主要指网络资源的描述数据，常用于网络信息资源的组织和利用【2】。元数据的目标主要有两个方面：一是简单高效的描述、保存、组织和管理大量信息资源；二是使信息资源的检索、发现、定位和共享更加便利与高效[3，4]。

2、元数据的结构

2．1 内容结构

用于定义元数据的构成元素。包括：描述性元素、技术性元素、管理性元素和复用元素等【5】。元数据内容一般分为三层，即元数据子集、实体和元素。元数据元素是元数据最基本的信息单元，实体是同类元数据元素的集合，子集是相互关联的元数据实体和元素的集合。在同一个子集中，实体可以有简单实体和复合实体两种，简单实体只包含元素，复合实体既包含简单实体又包含元素，同时复合实体与简单实体及构成这两种实体的元素之间具有继承关系【6】。

2．2 句法结构

用来定义元数据的格式结构以及如何描述这种结构。如元素结构描述方法(如XML Schema，RDF等)，结构语句描述语言(如扩展巴科斯范式标记法)等。此外，句法结构还可以定义元数据与被描述数据对象的捆绑方式。

2．3语义结构

用于定义元素的具体描述方法。包括元素本身有关属性的定义，一般采用ISO11179标准，《数据元素的规范和标准化》)；元素内容编码

规则定义，编码规则可以是特定标准，或最佳实践(best practice)或自定义的描述要求(in. smlc曲n)【7】。

3、元数据的类型

按组织信息资源的功能，元数据可分为以下类型：描述型元数据、结构型元数据、存取控制型元数据和评价型元数据。美国Getty信息研究所的Anne J. Gilliland—Swet—land根据元数据功能性将元数据划分为管理型元数据、描述型元数据、保存型元数据、技术型元数据和使用型元数据【7】。英国图书馆及信息网络化办公室将在结构和语义方面逐渐完善的一系列元数据分为三组：简单格式、结构化格式和复杂格式【3】。此外，还可以按照元数据的内部结构、应用领域、编码标记方式、开发设计角度、通讯协议等方面进行分类[8]。

4元数据特点和作用

4．1著录描述

元数据对数据单元进行详细、全面的描述。元数据元素包括内容、载体、位置与获取方式、制作与利用方法等方面信息。

4．2识别和确认

元数据对信息资源进行个性化描述，将信息资源中的重要信息抽出并加以组织，赋予语义并建立关系，提供识别和确认信息资源的基础，从而有利于用户识别和确认所需要的信息资源。

4．3评估与选择

根据元数据所提供的信息，参照相应的评估标准，结合使用环境和实际需要，用户可以对信息资源作出取舍，选择适合自己使用的资源。

4．4检索与定位

元数据通过在描述数据中提供检索点，实现对信息资源的检索和利用。由于元数据同时包含信息资源位置方面的信息，因此通过元数据可确定资源的存储位置，从而使用户获取所需要的信息资源。

4．5资源管理

元数据支持对资源利用和管理过程的政策与控制机制的描述，包括权利管理、电子签名、使用管理、支付审计等方面的信息。

4．6信息资源保护与保存

元数据支持对资源进行长期保存，包括详细的格式信息、制作信息、保护条件、转换方式、保存责任等内容。

4．7资源共享元数据可整合异质的信息资源，提供分布式信息资源共享[9，10]。

5、元数据的互操作

5．1元数据映射

所谓元数据互操作是指多个不同元数据格式的释读、转换和由多个元数据格式描述的数字化信息资源体系之间的透明检索。元数据映射指两个元数据格式间元素直接转换，或通过中介格式进行转换。这种途径转换准确、转换效率高，但在面对多种元数据格式并存的开放式环境中此法应用受到限制。

5．2标准描述方法

其过程是建立一个标准的资源描述框架，用这个框架来描述所有的元数据格式，只要系统能够解析这个标准框架，就能解读相应的元数据格式。通用标准是可扩展标记语言和资源描述框架，由于两者在应用中各有优缺点，所以有人提出将XML和RDF模式相结合的元数据互操作机制[11]。

5．3元数据复用

通过在一个元数据格式中，引用其它元数据格式的部分元素或属性，用来描述复杂资源，从而扩展元数据格式适用范围，以促进元数据的相互转换。

5．4元数据开放搜寻利用元数据来进行资源搜寻和发现时，可以采取元数据开放搜寻机制来实现元数据互操作。

5．5元数据语义转换

通过元数据语义定义和元数据概念集，支持两个元数据格式间元素通过语义分析进行转换。

5．6数字对象方式

通过建立包含元数据及其转换机制的数字对象，来解决元数据互操作。

6、元数据标准

6．1 定义

元数据标准是指描述某些特定类型信息资源的规则集合，一般包括语义层次上的著录规则和语法层次上的规定。主要用于数据发布、数据集编目、数据交换、网络查询服务等，同时也是数据集整理、建库、汇编、发布的依据。元数据标准一般具有适时、灵活、可扩展、易兼容和互操作性。

6．2结构

元数据标准体系整体上采用层次式的树状结构。即先由管理部门制定根级元数据标准，然后各学科领域根据根级元数据标准制定各自学科的领域元数据标准。由于相同层次上元数据标准的父标准都是统一的，

所以很容易对数据进行整合，并保证元数据的通用性，互操作性，也保证专业元数据标准可以为自己专业服务。通常一个元数据标准主要包括：前言、适用范围、参考标准、术语、元数据分级、元数据内容及定义、元数据扩展原则及相关附录等。

6．3制定原则

元数据标准的制定须遵从一定的原则，即标准要支持元数据在行业或其他领域的应用；标准以提供数据的轮廓为目的；标准要提供一个实体与元素集，并定义元素的性质(如必选，一定条件上可选及可选等)。元数据标准定义的对象是数据，而非定义与数据相关的计算机系统，传输手段和信息表现方式等。

6．4国内外应用的元数据标准

目前有几十种元数据标准，这些标准可简单划分为两大类：一类是针对互联网上信息资源的，典型的是都柏林核心元数据标准；另一类是针对行业的，如美国联邦地理数据委员会地理空间元数据内容标准等[4,6]。

7、元数据技术在科学数据共享中的应用

7．1元数据系统的组成

元数据系统主要由三部分组成：元数据编辑软件，用于编辑生成符合特定元数据内容标准规范的元数据文档；元数据库系统，由元数据库维护平台和元数据服务器两部分组成，用于元数据的管理、维护、网络发布；元数据网关，用于实现元数据的互联网发布，代理用户对多个元数据发布服务器的访问。

7.2元数据系统对科学数据共享的作用包括规范元数据;发布共享信息;促进元数据管理；减少重复生产和促进科学数据共享等。

7．3元数据共享体系的建立

首先，各科学数据提供机构建立自己的元数据系统，然后，将各参与机构的元数据系统的网关逐级联接，就形成了元数据共享体系。在这个体系中，网络用户可以通过任意一个元数据网关查询到所有在线元数据库中的元数据记录，从而实现元数据共享。最后，为实现科学数据的共享和交换，科学数据提供者可以根据自身条件和实际需要，建立相应的科学数据交易系统，以实现科学数据共享。

8、元数据在医学科学数据共享中的研究进展

8．1 国外

当前在医学领域有多种元数据方案并存，而这几种元数据方案都是基于Dc元数据标准。医学核心元数据方案(Med。ical Core Metadata，

MCM)。该方案是由美国完成，它是基于DC的医学核心元数据集，主要用于描述和组织网络医学信息资源。MCM继承DC的语义及语法结构。为适应对医学信息资源的组织和描述，MCM在DC基础上进行扩展。主要体现在两个方面：McM—MeSH Term表和MCM—Resource Type表。MCM—MeSH Term表采用MED—LINE数据库中的医学主题(MedicalSubjectHeadings，MeSH)；MCM—Resource Type表是为适合网络医学信息资源的新类型，在原出版物类型基础上扩展而来的，共有19项，其中新增类型有主页、文摘、新闻、图像、视频、音频、软件、病人教育和论坛等。此外,McM采用能够被网络系统识别和传递的超文本标记语言HTML予以标识，主要使用HTMI／Meta元素，用以标记关于资源整体的有关信息。法国医学资源元数据叫岫,是由法国Rouen University Hos. pital(RUH)发起，于1995年实施，主要用于描述和索引法语网络医学信息资源。法国医学资源元数据的设计思想与MCM基本一致，即基于DC元数据集和MeSH词表来组织医学信息资源。在医学资源类型描述控制方面，扩展MEDLINE出版类型列表，根据网络医学资源特性增加一些新的资源类型。EBM元数据方案的特点是：提供原始研究的结构式摘要；除MeSH主题词表以外，提出EBM实践中作为重要因素(研究类型、临床展望)的另外两个编码体系；提出EBM元数据用于因特网资源的可能性。它采用基于XML／RDF的可扩展标记语言，面向内容描述，文档结构灵活，自定义标记，增加了元数据的互操作性滔]。元数据及其技术在科学数据共享中的应用典范是“美联邦科学联盟元数据通道，该元数据整合中心集结各个学科领域的30个数据库和1 700多个科学网址，其中与医药卫生有关的数据库有5个，用户发一个检索指令，可以同时检索分布于美国的科学数据信息。其目的是为从事科学工作的公民及任何对科学有兴趣的公众，提供跨部门的检索通道来查找和使用政府提供的有关科学技术的信息资源。

8．2 国内

我国是一个国有科学数据大国，政府拥有的科学数据遍及科学研究的各个领域，其中在医学研究领域，已积累大量的基础、临床、预防和中医药方面的科学研究和观察数据，如中国医学科学院建立的基础医学数据库(包括中国人生理常数数据库、中国生物医学数据库、基因数据库等)和临床数据库(包括心血管病防治数据库、肿瘤数据库、高血压数据库和原发性骨质疏松症与老年性疾病数据库等)，其中基础医学数据库的数据量已达到10GB左右；中国疾病预防控制中心(CDC)建立的数十个同规模的预防医学数据库，其数据总量超过300GB；中国人民解放军总医院存有30多万份电子病历的电子病历数据库；中国中医研究院建

立的中药科技基础数据库、古代本草文献数据库、针灸文献数据库等多个数据库，其累积数据量已达到25GB，并且每年新增与更新数据约4G。随着医学发展，更多医学科学数据在不断产生，这些医学科学数据形成我国医学科学研究的重要科学资源。但是由于国内还没有制定专门针对医学科学数据共享的标准和构建医学科学数据共享体系，所以无法实现这些医学科学数据资源的检索、选择、交换、共享和有效利用。为实现医药卫生科学数据共享，充分发挥其应有的效益，2004年4月启动“医药卫生科学数据管理和共享服务系统”项目，其总体目标是建立基础医学、临床医学、公共卫生、中医药学、特种医学、药物与创新药物6个科学数据中心，并利用互联网技术将这6个科学数据中心连接起来，构成一个物理上分布、逻辑上统一的医药卫生科学数据管理和共享服务网。主要工作包括整合现有医学科学数据资源，建立50个主体数据库；制订数据共享规范及元数据标准；创建医药卫生科学数据目录查询系统等。现在国家信息中心正与相关单位协作制订有关科学数据共享工程的系列技术标准，包括：《科学数据共享元数据标准》、《科学数据共享概念与术语》、《数据模式描述规则和方法》、《元数据标准化基本原则和方法》、《元数据检索和提取协议》等。目前中国医学科学院、中国人民解放军总医院、中国中医研究院、中国疾病预防和控制中心等机构已联合制定《医药卫生科学数据共享元数据标准》。

参考文献

1. Cathro W，袁芳，苑二坡. 元数据研究概述[J]. 现代情报，2004，(4)：195—198.
2.马费成. 信息资源开发与管理[M]. 北京：电子工业出版社. 2004:122—129.
3.张敏，张晓林. 元数据的发展和相关格式[J]. 四川图书馆学报，2000，(2)：63—70.
4.冯项云，肖珑，等. 国外常用元数据标准比较研究[J]. 大学图书馆学报，2001，(4)：15—21.
5.张晓林. 元数据研究与应用[M]. 北京：北京图书馆出版社，2002:11—12.
6.王国复，吴增祥，臧海侍. 气象元数据标准与系统建设[A]. 见：孙九林，施慧中. 科学数据管理与共享[M]. 北京：中国科学技术出版社，2002:187—188.
7.秦笃烈. 从VHP看医学科学数据共享政策的重大意义和实施[A]. 见：

数字化可视人体国际研讨会论文集[C].Chongqing：International Workshop 011 Visible Human，2003:91—96.
8.李郎达．Metadata初探[J]．情报科学，2001，(6)：58—61.
9.刘嘉．元数据：理念与应用[J]．中国图书馆学报，2001，(5)：32—36.
10.高建勋，吴开华．元数据发展中的热点问题讨论[J]．图书馆，2002，(5)：41—44.
11.盛小平．元数据的互操作研究[J]．图书馆，2002，(2)：30—32．21
张大陆，刘畅．Web服务语义描述的架构[J]．计算机工程，2004，30(2)：73—75.