●吴开华 邢春晓 罗德胤

数字图书馆元数据研究

摘 要 元数据是数字图书馆建设的关键技术之一,是数字图书馆用以进行知识组织和资源发现的工具。它分为描述性元数据、管理元数据和结构元数据。随着越来越多的特殊科目和格式的元数据标准的出现,元数据的互操作问题已成为业内人士关注的焦点。参考文献 4。

关键词 数字图书馆 元数据
分类号 G250.76

ABSTRACT Metadata is one of the key technologies in the development of digital library, and is a tool for knowledge organization and resource development. In this paper, the authors analyze various kinds of metadata, and think that the interoperability of metadata will be a focus. 4 refs.

KEY WORDS Digital library. Metadata.
CLASS NUMBER G250.76

1 元数据的产生与发展

尽管元数据(Metadata)这个概念的提出只是近些年的事情,但这个概念的内涵所表达的信息却早已存在。其实,传统目录就是元数据的一种。其他如档案查找工具、博物馆的登记册等都是元数据应用的具体表现形式。因此可以说,无论是在数字化环境中,还是数字化环境之外,元数据都大量存在。

随着计算机技术的发展,传统目录逐渐被机读目录(MARC)所取代。MARC 无论是在数据描述的丰富性,还是在数据检索的查准率方面,都是其他元数据所不可比拟的。因此,在 20 世纪 60～90 年代,在以印刷型资料为收藏主体的传统图书馆里,MARC 独占鳌头,成为书目数据描述领域的主流工具。但是,由于 MARC 自身的局限性,例如结构复杂,著录项目过于烦琐,而且要求专业性强、要受过专门训练的编目员来做等等,虽保证了著录的精确性,但著录速度慢,工作效率低。面对网络海量信息资源的整序需求,MARC 无疑是满足不了的。因此,资源发现已成为 Internet 应用的瓶颈与焦点。虽说网络上已有多种搜索引擎,并辅之以布尔逻辑检索等方法,但同样不能满足用户对信息检索的需求,尤其不能满足用户对特定信息的准确检索。例如 Yahoo、Lycos、Altavista 等,这些搜索引擎的工作方式,是通过自动搜索程序来抓取网页信息,然后以自动拆字(词)做索引的方式建立数据库,不能有效地过滤资源,造成检索结果数量大而有用信息少的弊病。

元数据的产生为网络信息资源的组织提供了重要手段。所谓元数据,是用来标识、描述和定位网络电子资源的数据,是面向某种特定应用的机器可识别、可理解的信息。它去掉了 MARC 格式的烦琐和复杂,以结构化的字段检索弥补搜索引擎的不足,是介于 MARC 与搜索引擎之间的一种新型的数据格式。迄今为止,世界上已开发出并付诸使用的元数据有多种,例如:美国联邦地理数据委员会的地理元数据项目(FGDC, Federal Geographic Data Committee)、编码文档描述(EAD, Encoded Archival Description)、频道定义格式(CDF, Channel Definition Format)、教育管理系统(IMS, Instructional Management System)、全球信息定位服务(GILS, Global Information Locator Service)、博物馆信息计算机交换标准框架(CIMI, Computer Interchange of Museum Information)、互联网内容选择平台(PICS, Platform for Internet Content Selection)和都柏林核心元数据(DC, Dublin Core)等等。

已成为美国国家标准的都柏林核心元数据,是一个由 OCLC(联机图书馆中心)和 NCSA(美国超级计算机应用中心)在 1995 年 3 月联合发起,邀请来自图书馆、计算机、网络方面的学者和专家共同研讨的产物,目的是希望建立一套适合于网络电子资源的描述方法,以使得资源发现和信息检索变得更加迅速和有效。由于 DC 具有结构简单、语意互通和可扩展性等特点,因此,在众多的元数据种类中,成为最受瞩目的一种资源组织工具。

2　元数据的定义与内涵

通常,人们将元数据定义为“关于数据的数据”(data about other data)或“关于数据的结构化数据”(structured data)。在哈佛大学数字图书馆项目里,其定义为:元数据是帮助查找、存取、使用和管理信息资源的信息。在这个定义里,元数据既适合于电子资源、又适合于非电子资源;不仅包括编目信息,也包括其他管理和存取资源的信息。

元数据描述信息资源或数据对象,其目的在于使用户能够发现资源,识别资源,评价资源,而且对相关的信息资源进行选择、定位和调用,追踪资源在使用过程中的变化,实现信息资源的整合、有效管理和长期保存。

元数据的内涵包括语义、句法与内容标准。

语义定义了元素的含义。在对比研究两个元数据集时就可以发现相对应的元素。即如果一个元数据集里的 creator 和另一个元数据集的 author 都是指知识内容的主要创作者,那么我们就可以认为这两个元素是对等的元素,彼此可以互相映射。例如 CNMARC 中 200 字段的 $f 就可以与 DC 中的 creator 建立起互相映射关系。明确的语义定义是实现不同元数据互换的基础。

句法,是指句子的结构方式以及支配句子结构的规则。例如下面是 MARC、HTML、XML 三种不同句法结构的实例:

1001　$aGates, Henry Louis

〈META Name = “DC. Creator” CONTENT = “Gates, Henry Louis”〉

〈Creator〉Gates, Henry Louis 〈/ Creator〉

现在,我们暂且以 MARC 与 XML 两种不同句法结构作一比较:

1001　$aGates, Henry Louis 是用 US MARC 格式著录的某一特定记录的一个字段。在 US MARC 格式中,“100”的含义是:个人名称主要款目字段。记录用作主要款目标目的个人名称数据元素,一般为文献的主要个人责任者。紧接着的“1”,是该文字段的第一指示符,表明所记录的名称是以倒置形式著录,即姓在前,名在后。“1”后面是该字段的第二指示符,在这里是一个空位,表明该指示符未定义。随后的“$a”是该字段的子字段,其值是:个人名称。包含个人名称的姓、名、姓名、家庭名以及做主要款目标目的字母、短语、数字等。按照个人名称的形式,可将其划分为名、姓和家族名三种类型。根据以上关于 100 字段的字段、子字段以及指示符的含义与规定,我们知道,1001　$aGates, Henry Louis, 描述的是一个姓 Gates 名叫 Henry Louis 的某一特定文献的主要责任者。

〈Creator〉Gates, Henry Louis 〈/ Creator〉,是用可扩展标记语言 XML (eXtensible Markup Language)表示的某一特定文档的一个字段。〈Creator〉和〈/ Creator〉为控制标记,其中〈Creator〉为开始的控制标记,〈/ Creator〉为结束的控制标记。在小于号“<”和大于号“>”之间嵌套的是元素 Creator, 在〈Creator〉和〈/ Creator〉之间嵌套的是元素 Creator 的内容 Gates, Henry Louis。用元素表示组成文件的逻辑部件,元素概念明确,表达直观,并且用简单的嵌套和引用表示元素的内容。这种句法结构,较之 MARC 格式更容易理解,尤其方便计算机存储、处理、查询、传输、显示和打印。

内容标准之一是数据元素的格式。例如,有以下三种日期的表示方法:8. 6. 1999; 6/8/1999; 19990608,只有明确了数据元素的格式,我们才能知道它究竟是 1999 年 6 月 8 号,还是 1999 年 8 月 6 号。鉴于此,DC 发布机构建议其核心元素之一的 Date 格式应符合 ISO8601[W3CDTF]规范,并使用 YYYY-MM-DD 格式。那么,在 DC 的元数据标准中,1999 年 6 月 8 日应著录为:1999-06-08 才算符合规范。

内容标准的另一个子类是值标准。这包括分类表、主题词表等。例如某一具体的实用数据库,用的是 CLC(中国图书馆图书分类法)还是 LC(美国,国会图书馆图书分类法)的分类标准;是 CT(汉语主题词表)还是 LCSH(美国,国会图书馆的标题表)或者是 MESH(美国,医学主题词表)的主题词表,这在建库之初,就应该有一个明确的规定。元数据的值的标准的应用,是提高信息检索查准率和查全率的有效措施。

3　元数据的类型与作用

元数据有多种分类方式,例如按结构化程度可分为:非结构化元数据,如 yahoo 等搜索引擎;较少字段的结构化元数据,如 Dublin Core 等;高度结构化元数据,如 MARC 等。但最为普遍的分类方法是按照其功能分为描述性、管理和结构元数据。

描述性元数据支持资源的发现和鉴别。题名、作

者、制作者、出版者、出版日期等都是典型的描述性元数据。描述性元数据通常都是公共信息,因而它比别的元数据都得到了更好的标准的支持。MARC 和 AACR2 都是关于图书馆描述性元数据的标准。其他领域也有自己的标准,例如 Dublin Core 等。

用以维护和管理资源的数据通常被称为管理元数据。传统图书馆中有关借阅权限、馆藏地点等信息都是管理元数据的例子。管理元数据的范围很宽,一般包括:创建者元数据,用以表明谁拥有资源,谁承担资源存储的费用,谁有权改动甚至删除资源等;存取权限元数据,用来决定谁可以使用资源以及以何种方式使用资源;数据处理技术的元数据,如扫描分辨率和压缩率等,被用于将资源从一种格式转为另一种格式。

结构元数据描述数字化资源内部的形式特征,如目录、段落、章节等特征。它将资源的各个部分连接起来成为一个整体信息。将它用在程序里可以产生一个资源的显示界面,例如它可以将统计信息以图形的方式显示出来。结构元数据还可以包括支持在资源内部各个部分间浏览的信息,如翻动书页,用户可以从一页跳到另一页,从一个问题跳到另一个问题,从一本书跳到另一本书,从图像跳到和它相应的文本等等。

值得特别提出的是关于保存元数据的研究。保存元数据,指支持数字化资源长期保存的数据。在数字图书馆中,关于数字化资源发现与检索的元数据研究固然重要,关于支持资源保存的元数据研究同样非常重要。

所谓长期保存是指保存期限可以无限延长。关于保存元数据的某些研究内容,在描述性元数据和管理元数据的研究中也曾涉及到,只是将它单独列出并加以系统研究的却为数不多。不过,OAIS 是个例外,保存元数据的主要研究内容,OAIS 体现的最为充分。保存元数据的框架结构,OAIS 最引人瞩目。

OAIS 的全称是 Open Archive Information System。其主要功能是接收信息、保存信息并提供信息服务。OAIS 中的 PDI(Preservation Description Information),描述内容信息的特征以保证内容信息的长期保存,它包括下列内容:出处信息(provenance),描述内容信息的来源,产生以后的监管人、加工处理历史等;上下文信息(context),记录内容信息与信息包以外其他信息的联系;参考信息(reference),包括对资源描述的附加信息和资源标识符(用来标识内容信息的惟一性,例如一本书的 ISBN 号);固定信息(fixity),包括用于认证的信息,例如数字签名等,以保护内容信息不受篡改。

关于 OAIS,国际元数据界已逐渐达成共识。例如中国国家图书馆制订的《中文元数据标准方案》,就以 OAIS 作为总体框架;美国 OCLC/ RLG 已经正式提出保存元数据的概念并企图要在 OAIS 信息模型的基础上制订一个保存元数据的标准框架;澳大利亚国家图书馆也在保存元数据的研究方面做了许多努力。保存元数据这一概念的提出对数字化资源的长期保存具有深远意义。

4　元数据的互操作

随着数字图书馆的发展,从 90 年代开始,一般或特殊领域的元数据标准如雨后春笋般层出不穷,有些学者称之为"元数据运动"。这个"元数据运动"折射出"数据的结构数据"是对描述和发现各类资源属性的需求的一种本质和最初的反应。然而,随着越来越多的特殊科目和格式的元数据标准的出现,元数据的互操作问题已成为元数据的开发者和潜在使用者关注的焦点。

用户需要集成存储在异构数据库中的信息。但在过去的若干年中,不兼容的数据格式和数据结构阻碍了信息系统之间的互操作。一个日益重要的问题是怎样实现各种元数据间的互操作,并能够真正支持最广泛范围的资源发现、检索和使用。元数据的互操作直接影响信息的共享、互换以及透过系统、语言和地理位置的界限而访问的可能性。

目前,元数据的互操作主要是通过语义互操作和结构与语法的互操作来实现,有以下几种途径。

4.1　开放档案协议 OAI

OAI 的全名是 open archive initiative。OAI 的研究始于电子出版(E-Print)团体,最初的研究目标是通过电子出版团体内部系统的互操作来达到团体内的信息共享,后来将目标扩大为:寻求一种简便的方法来实现不同的数字资源系统间的开放检索(也就是跨系统检索)。目前它的具体做法是:以 Dublin Core 的 15 个核心元素为"中间件",使不同元数据方案下相等或近似相等的元数据元素相互映射,以实现语义上的互操作。

4.2　信息检索国际标准 Z39.50

Z39.50 是检索远程图书馆书目的信息检索国际

●纪蔚蔚　潘有能

文献数据库质量控制系统的实现 *

摘　要　文献数据库质量与建库成本、建库周期之间存在着矛盾。通过数据录改中的质量控制、主题词标引及分类中的质量控制、数据整合中的质量控制,可以解决这一问题。图 1。参考文献 2。

关键词　数据库　质量控制　文献处理
分类号　G250.74

ABSTRACT　There is a contradiction between the quality of a document database and the creation cycle and cost for the database. In this paper, the authors think that we can solve some problems by quality control in the inputting, subject indexing, classification and data integration. 5 figs. 2 refs.

KEY WORDS　Database. Quality control. Document processing.
CLASS NUMBER　G250.74

1　质量控制的提出

文献数据库是提供用户服务、制作光盘检索系统、进行因特网上查询和文献计量的基础,也是进行文献自动化处理的素材。数据库质量有问题,便会造成文献在数据库中的漏检或误检,还会造成自动标引的准确率低,甚至无法实现,文献计量指标及统计分析的可信度也不能得到保障。质量,是数据库的生命,它直接影响用户服务和文献利用,这已经成为人们的共识。

文献数据库的质量主要体现在:数据录入的准确性,数据收集的完整性,数据更新的及时性,数据处理的正确性。常见的错误类型有:字段类型不规范,相关字段不匹配,关

标准,是一个运行在 TCP/IP 协议之上的应用层协议,它规定了客户机查询服务器以及提取结果记录等过程中所涉及的数据结构和数据交换规则,从而解决了现存书目数据库检索接口的异构性问题。相对于 OAI,Z39.50 的功能更加完善,但也带来实现的困难和费用的高昂。一般说,只有标引详细,数据质量很高,对互操作质量要求相当苛刻的系统才采用。

4.3　资源描述框架 RDF 与可扩展标记语言 XML

由 W3C 推出的 RDF 是一套描述资源及其属性和属性值的模型,其制定的目的主要是为元数据在 Web 上的应用提供一个基础结构,以方便不同元数据间的互操作。可扩展标记语言 XML 作为元数据的编码标准,提供了元数据在语法层次上的互通性,使它跨越特定平台、特定系统的限制。使用 RDF/XML 命名域的概念,在创建一个元数据格式时,借用其他元数据集的某些元素,可以减少重复劳动并增强元数据格式间语义互通性,方便互操作的实现。

元数据是数字图书馆建设的关键技术之一。数字化图书馆的运作,无论是数据的加工、存取,信息的浏览、检索,还是资源的整合与长期保存都是以元数据为基础实现的。随着数字图书馆的发展,元数据的研究必将进一步深入。

参考文献

1　The Dublin Core Metadata Element Set. ANSI/NISO Z39.85-2001
2　Marcia Lei Zeng. Supporting metadata interoperability: trends and issues. Proceedings of 21st NIT international conference. Beijing: Tsinghua University Press, 2001
3　林海青. 数字化图书馆的元数据体系. 中国图书馆学报, 2000,26(4)
4　吴政猷. 都柏林核心集在图书馆应扮演的角色. 上海中文元数据应用国际研讨会,2001

吴开华　清华大学图书馆研究馆员。通讯地址:北京清华大学。邮编 100084。
邢春晓　清华大学计算机系副教授,博士后。通讯地址同上。
罗德胤　清华大学建筑学院博士研究生。通讯地址同上。

(来稿时间:2001-10-29)

*　本文系江苏公安专科学校科研项目《公安文献全文数据库及计算机辅助标引与检索系统》(97XB870001)的成果之一。