

●刘 嘉

元数据:理念与应用

摘 要 对元数据的定义有多种表述。元数据并不是新的概念,但在网络环境下有革新。元数据可分为管理、描述、保存、技术、使用等多种类型,在网络信息组织和检索方面具有重要作用。目前,元数据在世界范围内的应用越来越广;我国在数字图书馆领域里的应用已初见成效。表2。参考文献30。

关键词 元数据 定义 类型 作用 数字图书馆

分类号 G250.7

ABSTRACT There are different kinds of descriptions for the definition of metadata. Although it is not a new concept, it is innovative in the network environment. Metadata can be divided into several types, such as those for management, description, storage, technology and application. It plays an important role in the organization and retrieval of network information, and is being applied wider and wider. 2 tabs. 30 refs.

KEY WORDS Metadata. Definition. Type. Role. Digital library.

CLASS NUMBER G250.7

1 元数据概述

元数据的英文原文为 Metadata,最早出现于美国航空与宇宙航行局的《目录交换格式》(Directory of Interchange Format, DIF)手册中^[1]。在中国大陆,则将 Metadata 翻译为“元数据”;而在台湾则有“诠释资料”、“超资料”等多种译名。

1.1 何为元数据

元数据是描述和限定其它数据的数据。关于元数据,迄今为止,还没有完全统一的定义,最常规的定义就是:元数据是关于数据的数据(data about data)^[2]。一些专家和学者又把这个过于简洁的解释加以扩展和深化,较具代表性的几种定义有:

(1)元数据是关于数据的数据。此术语指任何用于帮助网络电子资源的识别、描述和定位的数据^[3]。

(2)元数据是关于数据的结构化的数据(structured data about data)^[4]。

(3)元数据是与对象相关的数据,此数据使其潜在的用户不必预先具备对这些对象的存在或特征的完整认识。它支持各种操作。用户可能是程序,也可能是人^[5]。

(4)元数据是对信息包(Information package)的编码描述(如,用 MARC 编码的 AACR2 记录、都柏林

核心记录、GIL S 记录等),其目的在于提供一个中间级别的描述,使得人们据此就可以做出选择,确定孰为其想要浏览或检索的信息包,而无需检索大量不相关的全文文本^[6]。

(5)元数据,即代表性的数据,通常被定义为数据之数据。它包含用于描述信息对象的内容和位置的数据元素集,促进了网络环境中信息对象的发展和检索^[7]。

1.2 对元数据的进一步认识

到目前为止,人们对元数据的认识还时常会有所混淆甚至产生谬误,笔者认为有必要就元数据做进一步的说明和解释。

(1)元数据不一定是数字形式的:元数据的形式是多样化的,并不一定是数字形式的。在管理人类文化遗产的过程中,有关专家一直在编制元数据。

(2)元数据不只同对信息对象的描述相关:专家们对元数据的描述或编目功能最为熟悉,但是,他们同样认识到元数据还能够说明被描述资源的使用环境、管理、加工、保存和使用等方面的情况。

(3)元数据可以来自各种不同的资源:元数据可以由人类(编制者、信息专家或使用者)提供,还可以由计算机自动生成,或者通过一项资源与另一项资源的关系来推断,如超链接。

(4)在信息对象或系统的生命过程中自然增加元数据:在信息对象或信息系统的整个生命过程中,

元数据首先被生成,继而被修改,甚或在某个阶段被废弃。

(5)关于元数据常规定义中的“数据”:在元数据的常规定义中,对前后出现的两个“数据”,应有较为宽泛的理解。一般而言,数据(Data)表示事物性质的符号,是进行各种统计、计算、科学研究、技术设计所依据的数值,或者说是数字化、公式化、代码化、图表化的信息^[8]。不过,在上述定义中,更准确地说,第一个定义所代表的含义应是“资源”。且两个“数据”都不一定是数字形式的,而可以以各种不同的形式出现。

1.3 元数据并不是全新的概念,但有革新

事实上,元数据并非什么新生事物。产生日久的图书馆书目记录就是一种元数据。但是,就联机信息对象而言,与其印刷品相比,技术已经带来了结构上的革新,从而使网络环境下的元数据发生革新。这些革新可以被分为新的单元、新的功能和新的模式^[9]。

(1)新的单元:新的单元源于生成数据对象的新技术,而非信息管理系统中所采用的新技术。新的元数据格式中所包含的许多单元,同适用于以往的印刷型信息资源的元数据格式中的单元相比,都必然是崭新的。

(2)新的功能:伴随着 HTML 的使用,目录记录、题名清单或其它地方中元数据单元的热链接为检索者生成了从外部制作的元数据到资源本身的直接的链接^[10]。通过使用相关性排序和概念映象之类技术的运算法则,搜索引擎可以模仿人工标引者的行为生成元数据。目录和数据库检索软件还能够被设定来隐藏元数据,不管是用于简要记录的显示还是自然语言的检索帮助。

(3)新的模式:计算机储存和网络信息资源检索的发展,已经引发了至少两种有关元数据的新模式的形成,即数据库模式(Database Paradigm)和标记模式(Markup Paradigm)。在数据库模式中,与数据对象相关的元数据被记录为以值来表示的属性;这些属性及其定义都被登记在一个数据词典里。每个元数据款目都是自治的,只取决于在数据词典中对它的定义和根据数据对象的含义所给定的标识符,而不依赖于其它的元数据。使用 MARC 格式的图书馆联机目录模仿了数据库模式。但严格地讲,缺少一个有效的数据词典,因为记录仍然是依据 ISBD 那种结构化的规则生成的,然后,被目录软件“转换”为一

列属性和值。由此而导致的冗余、不灵活、引起混淆的趋势在文献中常被论及^[11]。在标记模式中,涉及数据对象中元数据的嵌入,通常使用 SGML 文献类型定义来标识数据。或许,标记模式最大的潜力在于为数据本体内部的关系编码,用于基于人工智能的数据挖掘应用和精确的检索系统。

2 元数据的类型

Paul Miller 将元数据粗略分为两类:专家层次(Expert approach)与搜索引擎层次(Search engine approach)。前者包括 MARC、TEI 标题等较复杂的资源描述架构;后者则在 HTML 文件中隐藏<META>语法,使之可以被搜索引擎用于检索^[12]。显然,这种分类过于简单。

1998 年,美国 Getty 信息研究所(Getty Information Institute)曾就元数据进行过一次专项研究^[13]。在有关于此的专著中,Anne J. Gilliland - Swetland 根据功能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 5 种类型,其各自的定义及相应的例子如表 1 所示。

表 1 不同类型的元数据及其功能

类型	定义	例子
管理型元数据	在管理信息资源中利用的元数据	(1)采购信息 (2)权利和复制品追踪 (3)法定检索所要求的文献 (4)位置信息 (5)用于数字化的挑选标准 (6)版本控制
描述型元数据	用来描述或识别信息资源的元数据	(1)编目记录 (2)查找帮助 (3)特殊化的索引 (4)资源之间超链接的关系 (5)用户的注解
保存型元数据	与信息资源的保存管理相关的信息	(1)资源实体条件方面的文献 (2)保存资源的物理和数字版本中所采取行动,如数据更新和移植方面的文献
技术型元数据	与系统如何行使职责或数据如何发挥作用的元数据	(1)硬件和软件文献 (2)数字化信息,如格式、压缩比例、缩放比例常规 (3)系统反应次数的追踪 (4)真实性和安全性数据,如密码、口令
使用型元数据	与信息资源的等级和类型相关的元数据	(1)展览记录 (2)使用和用户追踪 (3)内容再利用和多个版本的信息

资料来源: Gilliland - Swetland, Anne J. Defining metadata.

In: Baca, Murtha, ed. Introduction to metadata: pathways to digital information. U. S. A.: Getty Information Institute, 1998. 3.

而 Lorcan Dempsey 与 Rachel Heery 则根据结构和语意的递增,将元数据分为以下 3 组,如表 2 所示。

表 2 根据结构和语意递增对元数据的分组

(全文索引)	(简单结构化的普通格式)	(结构更加复杂,特殊领域内的格式)	(更大的语义框架中的部分)
专属格式	专属格式 都柏林核心 IAFA/WHOIS ++ Template RFC 1807	FGDC MARC GILS	TEI 标题 ICPSR EAD CIMI

资源来源:Dempsey, Lorcan and Herry, Rachel. Metadata: a current view of practice and issues. The Journal of Documentation, vol. 54, no. 2 (March 1998): 156.

在表 2 中,第一组是取自资源本身的全文索引的数据,通常根本都不使用“元数据”这个术语。随着摘要和抽取技术的提高,其质量也会有所提高。因特网标引服务中使用的系统所生成的数据,即被包括在第一组内。第二组是简单的结构化的格式,使用范围相当普遍。它包括几个在计算机科学领域已经出现的支持检索和目录服务的格式。都柏林核心是此类格式中的一个例外,它是一个得到普遍认同的简单的元数据元素集。第三组是特殊领域内所使用的格式。在此,人们发现,更完整、更加结构化的格式通常都是为了满足特殊的功能需求和在特定的领域或医学传统内发展起来的。第三组中较新的格式通常都是基于 SGML 的,或正在向这个方向发展。第三组内某些格式的一个特征是,它们很明显的是包含了“内容”标识(如 TEI)的更宽泛的框架中的部分。当然,还有许多在表中未提及的格式^[14]。

3 元数据的作用

显而易见,元数据具有传统目录的“著录”功能,目的在于使资源的管理维护者及使用者可通过元数据了解并辨别资源,进而利用和管理资源,为由形式管理转向内容管理奠定必要的基础。

3.1 元数据在网络信息资源组织方面的作用

3.1.1 描述(Description)

根据元数据的定义,它最基本的功能就在于对信息对象的内容和位置进行描述,从而为信息对象

的存取与利用奠定必要的基础。

3.1.2 定位(Location)

由于网络信息资源没有具体的实体存在,因此,明确它的定位至关重要。

元数据包含有关网络信息资源位置方面的信息,因而由此便可确定资源的位置之所在,促进了网络环境中信息对象的发现和检索。此外,在信息对象的元数据确定以后,信息对象在数据库或其它集合体中的位置也就确定了。这是定位的另一层含义。

3.1.3 搜寻(Discovery)

元数据提供搜寻的基础,在著录的过程中,将信息对象中的重要信息抽出并加以组织,赋予语意,并建立关系,将使检索结果更加准确,从而有利于用户识别资源的价值,发现其真正需要的资源。

3.1.4 评估(Evaluation)

元数据提供有关信息对象的名称、内容、年代、格式、制作者等基本属性,使用户在无需浏览信息对象本身的情况下,就能够对信息对象具备基本了解和认识,参照有关标准,即可对其价值进行必要的评估,作为存取与利用的参考。

3.1.5 选择(Selection)

根据元数据所提供的描述信息,参照相应的评估标准,结合使用环境,用户便能够做出对信息对象取舍的决定,选择适合用户使用的资源。

3.2 元数据在网络信息检索中的作用

3.2.1 管理大量低网络带宽的数据

元数据致力于解决标引大量不同形式的数据而无需数量庞大的网络带宽的问题。标引得到的是代表性数据,而非信息对象本身。

3.2.2 支持有效的网络信息资源的发现和检索

当元数据单元和结构被设计来在深度上分析数据的内容时,它就促进了信息的更复杂和更综合的检索。

3.2.3 分享和集成异构的信息资源

信息资源以不同的形式具有不同的特征存在于异构的数据库中。标准的元数据描述允许在分散的网络环境下比较、分享、集成和再利用不同类型的数据。如此,元数据就成了在异构数据库中找到信息的重要的长途径。

3.2.4 控制限定检索的信息

元数据不仅能够促进有效的异构信息资源的检索,还能够管理限定检索的信息和客户服务,如排

序、过滤和评分、保密性和安全性。元数据起到了看门人的作用,具有永远增长的商业化信息资源所不可缺少的特性^[15]。

此外,从系统的角度审视元数据,元数据的功能还包括提供浏览及检索的功能、管理功能以及组合各个对象以及藏品的再呈现等^[16]。

4 元数据的应用

元数据的使用范围非常广,Susanne Boll 等认为元数据可被应用于图像检索、导航和图像集中的浏览,视频、音频和演讲,结构化的文献管理,地理和环境信息系统,数字图书馆,支持信息存取的混合多种媒体等^[17]。

4.1 元数据在整个世界范围内的应用

在整个世界范围内,元数据正在被用于越来越多的领域中。比较突出的实例即有:

(1)教育资料网关(Gateway to Educational Materials, GEM)^[18];

(2)欧洲数学图书馆和电子资源(The EULER project, European Libraries and Electronic Resources in Mathematical Sciences)^[19];

(3)北欧万维网索引(Nordic Web Index, NWI)^[20];

(4)亚历山大数字图书馆(Alexandria Digital Library)^[21];

(5)DESIRE (Development of a European Service for Information on Research and Education,欧洲研究与教育信息服务之发展)项目^[22];

(6)北欧元数据项目(The Nordic Metadata project)^[23]。

在元数据的众多应用领域中,数字图书馆非常突出。目前,所有正在开发中的数字图书馆项目都必须解决元数据方面的问题。这些项目中的大部分已经使它们的元数据框架能够适用于数字数据馆藏的描述,这些资料不仅包括全文数据,还包括参考资料集以及视频和音频那样的多媒体资料等。在数字图书馆中,元数据的主要作用是为分布式资源发现和检索奠定基础,元数据体系具备描述、整合、控制和代理四个基本功能^[24]。

此外,元数据在知识管理中同样能够发挥重要的作用。因特网及万维网的快速发展,促成了对于网络空间中数字信息和知识的传递和获取。有鉴于此,知识管理成为当今的重要课题。元数据能够通

过完整描述揭示对象的内涵,自然能够在知识管理中发挥作用。元数据贯穿于信息对象的整个生命过程,渗透于知识管理的各个环节。知识从数据到信息的增值过程中所采取的方式,主要与信息对象的形式特征有关,最根本的依据是关于信息对象的元数据;从信息到知识的增值过程,则不仅限于外部的信息加工与组织,而需要对信息对象内容的加工和分析,其根本依据仍然是元数据。在知识管理的前期阶段—信息增值(资料—信息—知识)的阶段中,元数据发挥着重要的作用;而在知识管理中的其它阶段中,元数据同样不可或缺。

4.2 元数据在中国的应用

在台湾,元数据的应用项目主要有由台湾辅仁大学的吴政睿教授开发的分散式元资料系统和梵蒂冈地区中文联合馆藏系统以及1998年已开始实施的“数位博物馆”(Digital museum)项目等,后者包括“资源组织与检索之规范”(Resources Organization and Searching Specification,简称ROSS)等多个分项计划^[25]。而在中国大陆,迄今为止,最成熟的元数据标准是中国机读目录格式。除此之外,其它元数据在中国大陆的应用与发展尚处于起步阶段,其中,比较突出和初显成效的是数字图书馆中元数据的应用。

目前,在中国大陆,影响最大的数字图书馆有:

(1)中国试验型数字式图书馆:中国试验型数字式图书馆项目(China pilot Digital Library project, CPDLP),是1997年度由国家计委批准立项的国家重点科技项目,项目实施期限为1997年7月至2000年12月^[26]。项目组规定该数字图书馆项目最小元数据集采用都柏林核心;最小元数据集都柏林核心的实现采用RDF方式;元数据分成两层,第一层为都柏林核心,第二层为MARC或TEI标题,在资源建设中要求按这两层实现。技术组据此提出项目所需采用的规范的元数据元素集及其定义、元数据元素的限定规则以及元数据元素的具体表达方式^[27]。

(2)中国数字图书馆工程:1998年8月25日,“中国数字图书馆工程”筹备领导小组成立,有关研究、论证工作正式展开^[28]。在这个庞大的项目中,元数据方案的制订是最基本和至关重要的部分,需要建立数字图书馆国家中心的元数据共享、检索系统,建立一个元数据资源中心。该中心使用并行数据库技术和分布式计算机系统来支撑海量的元数据系

统。

(3)清华大学建筑数字图书馆:清华大学建筑数字图书馆(THADL, Tsinghua University Architecture Digital Library)是由清华大学图书馆、计算机科学与技术系、清华大学建筑学院三方精诚合作共同研制开发的。日前,在清华大学数字图书馆站点上,已能够提供中国营造社史信息导航、学社资料成果信息导航以及新营造学社资料方面的服务。清华大学数字图书馆的元数据基本上采取的是都柏林核心^[29]。

此外,中国大陆有关方面正在着手于基于中文资源的元数据标准的制订。广东省中山图书馆制订的《数字式中文全文文献通用格式》^[30]是这方面的尝试和努力结果之一。北京大学数字图书馆研究所、北京大学图书馆数字图书馆工程将“中文 metadata 标准研究”作为重点研究项目。

目前,中国大陆数字图书馆的针对对象大都偏重于古籍、珍贵史料。随着数字图书馆在中国大陆的发展,数字图书馆的对象将逐渐丰富、扩大。

参考文献

- 1 NASA, Directory of Interchange Format Manual. Versaion 1.0. July 13 1998. NSSDC/ WDC - A - R &S. 88~89.
- 2 Hannemyr, Gisle. The internet is not your friend: using the Dublin Core on the Internet. Internet WWW page, at URL : < http :// www. ddb. de/ partner/ dc 7 conference/ results/ dc 7 - hannemyr/ sld001. htm > (version current at 22 May 2000) .
- 3, 15, International Federation of Library Associations. Digital libraries: metadata resources. Latest revision at September 29, 1999. Internet WWW page, at URL : < http :// www. ifla. org/ II/ metadata. htm > (version current at 12 March 2000) .
- 4 Iannella, Renato. Mostly metadata: a bit smarter technology. Internet WWW page, at URL : < http :// www. dstc. edu. au/ RDU/ reports/ VALA1998 > .
- 5, 14 Dempsey, Lorcan and Herry, Rachel. Metadata: a current view of practice and issues. The Journal of Documentation, vol. 54, no. 2 (March 1998) : 149, 157.
- 6 Taylor, Arlene G. The organization of information. Englewood, Colorado: Libraries Unlimited, Inc., 1999. 246.
- 7, 16 Xu, Amanda. Metadata conversion and the library OPAC. The Serials Librarian, vol. 33, no. 1/2, 1998: 180, 181.
- 8 周文骏主编. 图书馆学情报学词典. 北京: 书目文献出版社, 1991: 403
- 9 Wool, Gregory. A meditation on metadata. The Serials Librarian, vol. 33, no. 1/2 (1998) : 170~172.
- 10 W3C. Hypertext Markup Language (HTML) Internet WWW page, at URL : < http :// www. w3. org pub/ WWW/ Markup/ Mark Up. html > (version current at 13 March 2000) .
- 11 Leazer, Gregory H. An examination of data elements for bibliographic description: toward a conceptual schema for the USMARC formats. Library Resources & Technical Services, 3 (Apr. 1992) : 189~208.
- 12 Miller, Paul. Metadata for the masses. Internet WWW page, at URL : < http :// www. ukoln. ac. uk/ ariadne/ issue5/ metadata - masses > .
- 13 Gilliland - Svetlan, Anne J. Defining metadata. In: Baca, Murtha, ed. Introduction to metadata: pathways to digital information. U. S. A. : Getty Information Institute. 1998. 6~8.
- 17 陈雪华, 陈昭珍, 陈光华. 数字图书馆/ 博物馆中诠释资料之理论与实物. 图书馆学刊, 第 13 期 (1998 年 12 月) : 41~42
- 18 Sheth, Amit and Klas, Wolfgang, ed. Multimedia data management: using metadata to integrate and apply digital media. New York: Mc Graw - Hill Co., Inc., 1998. 5~10.
- 19 Sutton, S. A. ; Oh, S. G. GEM: using metadata to enhance Internet retrieval by K~12 teachers. Bulletin of the American Society for Information Science, vol. 24, no. 1 (Oct/ Nov 97) : 21~24.
- 20 Jost, M. EULER - Ein EU - Projekt zur Integration heterogener Informationsquellen. Bibliotheksdienst, 32 (3) (1998) . 513~517.
- 21 Ardo, A. ; Lundberg, S. A regional distributed WWW search and indexing service - the DESIRE way. Computer Networks and ISDN Systems, 30 (1/7) (Apr 1998) : 173~183.
- 22 von Essen, F. Metadata Initiative of German Libraries (META - LIB) = Metadaten - Projekt deutscher Bibliotheken: [a project]. 1997~1999.
- 23 Day, Michael. Metadata: mapping between metadata formats. Last updated: 02 Feb. 1999. Internet WWW page, at URL : < http :// www. ukoln. ac. uk/ metadata/ interoperability/ > .
- 24 The Nordic Metadata Project. At URL : < http :// Linnea. helsinki. fi/ meta/ index. htm > .
- 25 林海青. 数字化图书馆的元数据体系. 中国图书馆学报, 2000(4) 59~69

(下转第 45 页)

表2 模数转换的程序与问题

步 骤	涉及问题
拟数字化对象的选择 ↓	文献价值;版权;原件状况与完整性;数据丢失后,原件可否再扫描;文献被数字化的范围(局部或全部)。
按需组织信息内容 方案制定(如文档 页面图像或可检 索的文本数据库) ↓	适当的文件格式与存储媒体选择。
数字化前的 原件准备 ↓	保留原件吗?(建议保留)原件完整 性与强度检查。
品质检查 ↓	需要具有可接受的长期保存质量要求;如果由第3方提供技术服务,是否有文档说明与质量担保。
最终的索引与目标 ↓	资源开发的元数据,管理元数据与数字拷贝的保存。
将数据装填进 计算机 ↓	制作适当的备份拷贝。建议制作4个拷贝,至少有2个是分开保存在不同地理位置。
执行归档与保存 策略,或传送到 保存部门	格式、存储媒体、文档、转换程序均应标准化;母文件与备份的存储;媒体更新以及对付技术淘汰的策略。

字化过程中,还必须考虑存取政策与控制措施、费用

机制与付帐程序等。受版权保护的文献,可使用口令、直接付费、权限控制等措施控制使用;对不必控制使用的回溯文献数字化产品,则可以对阅读、下载与打印以不同价格对不同用户以不同规定加以管理。

在管理问题上还必须考虑,模数转换的费用是否可与取得的效益相平衡,以证明该项馆藏的数字化是合理的。费用/效益评价是十分困难的,但又是极为必要的。据有关资料介绍:十余年来,国家已投入10亿元制作了近千个数据库,但据统计,真正有市场价值的只百余个,能提供商品服务的数据库不到10%。大多数是年年靠财政上的不断输血来维持^[3]。

以上仅仅是从宏观上对馆藏文献数字工程加以考虑,在实际操作中,需要考虑的问题更多更细,表2是数字化处理中必须考虑的有关问题。

由上可见,馆藏文献数字化工程是十分复杂的。数字文献形成后,是否可以达到预期目的,既取决于拟数字化对象的选择,也取决于数字化处理过程的质量控制。

参考文献

1,2 <http://www.thames.rlg.org/preserve/matrix.hemL>
3 董焱,刘兹恒.图书馆馆藏文献数字化.图书情报工作, 2000(7)

刘家真 武汉大学信息管理学院教授,博士生导师。
通讯地址:武汉市。邮编 430072。

(来稿时间 2001-05-08)

(上接第 36 页)

26 刘嘉.元数据之研究.北京大学信息管理系博士论文, 2000:176~179
27 Zhuang Jing.中国试验型数字式图书馆项目. 1 Oct. 1998. Internet WWW page, at URL : < <http://www.nlc.gov.cn/dlib/dcl.htm> > (version current at 15 May 2000)
28 刘炜,赵亮.元数据方案实施意见. 15 June 1998. Internet WWW page, at URL : < <http://www.istis.sh.cn/istis/dlib/report/metadata.html> > (version current at 25 January 2000).
29 中国数字图书馆工程筹备领导小组办公室.“中国数字图书馆工程”筹备领导小组成立,有关研究、论证工作正

式展开.演示文稿,第1期(1998年8月26日). Internet WWW page , at URL : < <http://www.ccnt.gov.cn/digilib/dilijbl.htm> > (version current at 16 May 2000).
30 清华大学建筑数字图书馆项目组.清华大学建筑数字图书馆. Internet WWW page , at URL : < <http://166.111.68.208/digital/default.asp> > (version current at 7 January 2001).
31 广东省中山图书馆.数字式中文全文文献通用格式:征求意见稿. 2000-01-22

刘嘉 博士,北京大学信息管理系讲师。通讯地址:北京大学。邮编 100871。

(来稿时间:2001-04-16)