

●曾 蕾 张 甲 张晓林

元数据标准的演变

摘 要 网络时代,元数据有以下特征:功能上,描述从单一到综合;结构上,从整体式到模块化,从万变不离其宗的大而全到千变万化的随机组合;形式上,从繁到简,并可以自由伸缩;覆盖面,从包罗万象到面向特定学科、专业、文献类型;表现方式,从数字型字段代码到文字型含语义的元素标签,并逐渐实现 XML 化;生成方式由手工向自动化方向发展。参考文献 2。

关键词 数字图书馆 元数据 元数据标准 网络

分类号 G254.36

ABSTRACT In the age of network, metadata have the following characteristics: from singleness to integration in functions, from comprehensiveness to modularity and from all-inclusiveness to random combination in structures, from simpleness to flexible complexity in forms, from all-inclusiveness to subject-orientation in coverage, from numerical field tags to semantic character-type element tags and then to XML in representation methods, and from manual work to automation in data generation. 2 refs.

KEY WORDS Digital library. Metadata. Metadata standards. Network.

CLASS NUMBER G254.36

最近 10 余年来,随着网络化数据交流的普及应用和数字图书馆的崛起,元数据的重要性受到广泛认识,与其相关的标准的发展日益引人注目。元数据的概念部分起源于图书馆编目工作,但是它的发展趋势又超出图书馆的机读目录格式的格式标准。如果不理清两者之间的关系,我们就无法在元数据标准的研究与应用上取得快速的进展。

1 怎样看待前网络时代机读目录格式

机读目录格式是图书馆编目规则与计算机磁带存储格式的结合。图书馆编目规则是长期图书情报工作的经验总结,它既是元数据概念的渊源,也是目前处理书目数据、产生描述元数据的最佳方式。计算机磁带存储格式是 60 年代计算机技术的产物,也是目前机读目录格式继续发展的主要障碍。机读目录格式的目录区、指示符和子字段都是为适应顺序存取的技术要求而设置的。据此,我们可以说机读目录格式是元数据标准中的一种,它需要改进的部分是数据的组织结构,也就是如何将现有的数据元素根据新技术标准(如可扩充置标语言 XML)组织起来。

为了发挥机读目录格式的作用,而且弥补现在元数据中缺乏比较详细的综合性的描述元数据的缺陷,机读目录格式已在美国国会图书馆的标准工作

组的操作下,脱胎换骨,先是从机读目录格式中衍生出来的元数据物件描述格式表 MODS(Metadata Object Description Schema),紧接着是机读目录格式的 XML 表现形式(MARC XML),再加上元数据编码与传输标准 METS(Metadata Encoding and Transmission Standard)的完善,使机读目录格式及任何其他描述元数据记录都能被融入一个综合的元数据记录,“古老”的机读目录格式终于又受到人们的青睐,其魅力在新的一代数字图书馆中逐渐发挥出来。

在网络时代,如果简单地将机读目录格式划到元数据标准之外,排斥机读目录,我们就无法将书目记录与数字图书馆结合在一起,会将目前仍是主要信息源的书目记录排除在数字图书馆之外。但是,若一味坚持采用机读目录格式而阻止新的元数据标准开发则会影响整个数字图书馆工作的进程,这是我们需要避免的另一个极端。

2 网络时代数据类型的特点

元数据是关于数据的数据,换句话说,元数据需要容纳可能出现的任何类型的数据。多样化的数据类型包括书目型数据、文献型数据(半结构化或无结构)、数字型数据(图、声、像)、数值型数据(地理数据、统计数据、软硬件数据交换)等。针对这些数据而产生的元数据产品远远超出了机读目录格式时代“编

目'概念的内涵。与处理单一数据类型的机读目录格式相比,现今元数据所要担负的任务超过传统图书馆过去能承担的范围,涉及到我们生活的每个方面。从数量上看,现在一年中被处理和将被处理的数据比过去图书馆处理了几十年的数据还要多。

3 网络时代元数据的特征

数据类型的多样化和数据量的猛增、数字化信息用户群体的形成以及信息技术的迅速发展,促成了元数据从量变到质变的飞跃。我们以下分别从几方面来考察。

3.1 功能:从单一(描述)到综合(描述、管理、技术、保护等)

机读目录格式的确应被纳入元数据标准或元数据集的范围,但它只是所有元数据中一个侧重于书目描述用的元数据标准,描述在这里基本上是其唯一的宗旨。机读目录格式基本上满足了对各类型出版物描述的需要,例如早中期的机读目录格式分别为单本书、连续出版物、地图、缩微资料、音像资料等等编制了特定的格式。到后期这些形式被有机地统一起来了,但是其基本作用和使用方式并没有变。以机读目录格式为代表的数据库记录是在非数字化、非网络化时代为了详细描述图书馆所收藏的出版物而设计的,比起卡片式和书本式书目记录来,机读目录格式机读记录已经是一个飞跃。自从有了联机目录,起码读者不用到几十里外或几百里外去查目录,或去验证某一本书或文献到底是不是自己所要的。机读目录格式将过去印在卡片和书本目录上的信息用机器可读形式存储在数据库中,不仅方便管理、查询,而且通过联机形式得到信息共享,大大提高了图书馆工作效率。但是,这种工作程序的主要功能局限在“描述”方面,一条机读目录格式数据往往要动用几十个字段和子字段,将一本书或其他实体的物理特征和内容特征详细地记录在编码中。这种详细程度是因为书目记录与图书馆馆藏中的实物是分开的,书目记录需要帮助读者根据目录上的描述决定此书是否就是他所要的。从目录厅到书库,又从一个目录柜到一排书架,最起码也要有点距离,一个读者根据目录上的描述,决定一本书是否是他所要的。在开架的情况下这些书可以直接由读者从书库中把书取出来。在闭架的情况下,书单被送进书库,由馆员取来(至今在美国国会图书馆、纽约公共图书馆等有世界一流藏书的图书馆仍是以闭架取书为主)。

这种情况下,当要靠书目记录来决定取舍时,人们当然总是希望对实体的描述越彻底越好。

数字化馆藏、特藏和数字化图书馆的基本特点之一是信息源的数字化,几乎所有形式的出版物和其他信息源都可以用数字化的形式来表现和存贮。这几年,进展最明显的主要是图像、声音文字材料等的处理。一个人在计算机目录上找到一条关于某作品的记录时,就可以直接点击联到原作的数字化版本。这种情况下,对元数据记录关于文献源的详细描述的要求就自然而然地降低了。另一方面,由于一个作品现在往往不仅有其最原始的出版形式,通过数字化等过程又产生一个或多个衍生品,这就要求在书目数据中加入管理元数据内容,阐明每个原作和衍生作品的版权所有、使用权限、参与部分制作的单位的情况等。同时,每个数字化的文件在使用中又有对平台、软件、插入软件(PLUGIN)等的要求,有在其数字化过程中经过处理的各种参数,例如存贮格式(PDF,DOC,JPG,GIF等)扫描清晰度、体积、文件SIZE等。这些则要求在书目数据中加入技术元数据内容。根据各种类型数字图书馆的需要,有些记录中还要求加入保存元数据内容。这样,现在的元数据的功能不再仅仅是描述或以描述为主,元数据集描述、管理、结构、技术、保存等信息为一体,在数字图书馆中的功能早已超出了单纯的书目描述功能。

3.2 结构:从整体式到模块化

描述、管理、结构技术、保存等元数据元素的存在给机读目录格式时代的元数据带来了另一些特点。首先,元数据可以是模块化的,即描述、管理、结构技术、保存等方面的元数据可以各自在自己的模块中形成,有些是批量形成,有些可以是自动形成。模块不一定要由同一个人、同一个部门做成,而形成的模块又可以被反复使用、分散使用或修改,既避免重复,又保证了一致性。例如,一幅关于武汉黄鹤楼全貌的摄影作品经数字化处理后可能产生好几份文件,例如按不同清晰度扫描或存贮成TIFF或JPG格式,这几份文件的元数据记录中,描述元数据部分可能基本上是一样的,但是技术元数据部分则不同。又如,该摄影者将图片的出版印刷等权利卖给了不同的出版者,其中有一处将照片放入网上教学教材中,另一处将其加工制成电子明信片的形式,还有一处将其电子版复原成与原件一模一样的作品,存在档案中,请注意,这时的电子版文件是由不同的单位

制作的。这种情况下,该照片的描述部分可能还是相同的,但其管理和技术元数据成分则各有不同。

模块化的另一个好处是不要求采用唯一的、统一的元数据格式标准。大家不需要非要采用机读目录格式或都柏林核心集的格式来生产元数据记录。有必要时,不同格式下产生的数据可以互相转换,比如在甲方可以将乙方的原始元数据记录根据甲方需要转换过来。有的时候,这种转换甚至都成了多余的。例如 METS 的思想,是像粘胶一样,把不同模块粘在一起,人们可以把描述部分放上详细的数据,也可以只是放上引导(链接)到该描述元数据记录的 URL。对结构、技术、管理等部分也可照此处理。这种粘胶的方法,原则上允许任何形式的原始数据引入。

模块化改变了过去用统一一种格式处理所有文件、万变不离其宗的大而全方式、全国乃至全世界“统一”化的方式。机读目录格式时代允许千变万化的有机组合的局面,对于充分利用各种原始数据特别是多年累积的书目数据,来兴建数字化馆藏和数字化图书馆提供了先决条件。

3.3 形式:从繁(几十个字段,上百个字段)到简(10 来个核心元素),并且可自由伸缩

机读目录格式时代的元数据也以简略著称。可以说,“简单化”和“实用性”是一些主要元数据标准的基本准则。过去机读目录格式本身的烦琐,加上对英美编目条例 AACRII 的严格实施,和对国际标准书目描述格式(ISBD)的标点符号的严格采纳等,使许多非图书馆编目人员望而生畏。比如,编目人员除了要了解机读目录格式的固定字段和非固定字段、可重复和不可重复字段的用法,还要掌握成百个字段的场合、第一和第二个指示符(INDICATOR)的取值条件等。每条机读目录格式数据的产生往往需要耗费相当多的人力物力,而且只有经过严格训练的编目人员才能做出符合质量要求的书目记录来。对机读目录格式编目烦琐哲学的挑战首先是都柏林核心集(Dublin Core, DC)的“打群众战争”的思想。面对浩瀚的网上资源、与日俱增的电子化资源,过去那条机读目录格式编目的道路是行不通的。DC 工作小组最早提出的,就是要让每个网上资源的作者本人成为其作品的编目者,其采用的“核心”一词,一语道破天机——不要烦琐的面面俱到、层层区分的格式,而只要能揭示最基本特征的核心元素。经过几次改进,DC 形成 15 个核心元素的一

套定义,根据需要,每个元素都可以重复使用,也可以被弃之不用,十分灵活。

简单化并非是唯一的原则,它的实现还要靠可伸缩的原则。都柏林核心集的可自由伸缩性保证了某些具体项目的需要,例如,对需要具体一些的数据源,DC 又定义了一套“特征词汇”(QUALIFIER),允许在一个元素下按照特征进行细分;此外,对于面向类用户、资源的数据源,允许在 DC 基础上增加少数元素(例如 DC-教育版)。尽管 DC 从纵向(扩展的“特征”)和横向(增加的专用元素)都有所发展,在申报国家标准时,仍仅有 15 个核心元素被纳入和定为国家(NISO)标准。在利用 DC 的基础而开展的 OAI(开放描述数据)的基本协议中,也只有 15 个核心元素被 OAI 接纳,任何要参与在 OAI 下进行的数据交换和投放库的数字图书馆,在自己的资源中虽然仍可保留比 DC 更详细的记录,在交换中则必须将数据转换输出或允许匹配为与 DC15 个元素一致的格式。

在机读目录格式之后制定的元数据标准,有简有繁,一般来说越早期的越烦琐,比如档案元数据格式(EAD)是继机读目录格式之后为了解决档案材料的处理问题而编制的,其目的是协助查找档案(FINDING AIDS),因为档案资料是以盒、文件夹为单位的,FINDING AIDS 也是为了检索者通过这个目录来了解档案资源。虽然 EAD 及时采用了标准通用置标语言 SGML 为表现形式,其内容仍十分烦琐。由 GETTY 出面组织制作的描述艺术作品的元数据范畴(Categories for the Description of Works of Art, CDWA)也是较早问世,可以说十分详细(好几十个元素)。与其对应的是后来居上的图像资源协会的原数据核心集 VRA CORE,其对“核心”(Core)一词的采纳表明了它走简单化道路的意向。VRA Core 最早将描述作品(Work)本身的元素与描述作品图像(image)的元素分成两大组,共有 30 多个元素,后来将二者合并,形成今天(第 3 版)的 17 个核心元素类目。

3.4 覆盖面:从包罗万象到面向特定学科、专业、文献(物件)类型

在欧美国家虽说没有“百花齐放”、“百家争鸣”一说,机读目录格式时代的元数据发展却正是这种精神的体现。多少年来,机读目录格式是处于垄断地位的唯一描述格式,大大小小的图书馆,不论什么专业,都以机读目录格式为自己的编目格式。机读目录格式时代,这种局面已经一去不复返,许多以

某个特定学科、专业(如地理、艺术)、某个用户群(如教育)、某种文献源形式(如图像资料、毕业论文)、某种用途(如拍卖、鉴定)等为目的而编制的元数据标准应运而生,如果把元数据集、数据字典和以可扩展置标语言(XML)的格式表(Schema)或文件类型定义(DTD)形式发表的数据格式与定义都纳入人们所说的“元数据运动”,其覆盖面和种类的确可以用“如雨后春笋、层出不穷”来形容。而且,这些面向专门使用对象的标准格式可能互相之间有所重复,例如在处理艺术作品时,人们可以看到起码5种比较通用的标准格式。它们可能是由有不同背景的人或团体制作的,发表元数据标准的团组有计算机行业、非营利机构、学会、协会等等。

3.5 表现方式:从数字型字段代码到文字型含语义的元素标签,并逐渐实现XML化

机读目录格式时代的元数据几乎无一例外地采用了带语义的文字标记,虽然有少数元数据标准(例如IEEE LOM和DLSE)的标记仍然带有数字,其作用只是将元素归类,将相关的元素集合在一起,实际应用中的数据仍是文字标记。近两年来XML的成熟和普遍应用,以及XML编辑工具不断完善,大大促进了元数据格式的XML化趋势。有些元数据标准在发表之时既以XML的格式表或文件类型定义形式出现(例如档案元数据),更多的是在将元数据及元素间关系严格定义后,另将其用XML形式发表(例如LOM、MODS)。当然有的时候,XML版本要晚于原定义的发表。例如LOM的语义定义经过不断修改已经公布了第6.3版,但其XML DTD仍反映的是2.1版的内容。MODS在这方面做得好一些,它集中了分散在机读目录中的相关字段(例如100字段和700字段都是关于作者的),又用带有清晰语义的文字形式作为元素标记(TAG),不再用数字作为标记。MODS公布时,其XML Schema同时问世。

最有轰动效应的是MARC-21-XML征求意见稿的问世,这项工程不仅将机读目录格式元素全部用XML SCHEMA形式表现,而且提供了相应的将传统机读目录格式记录转换成MARC XML记录的工具,这项工作将大大促进图书馆将书目数据XML化、文字化,对图书馆数字化特藏或数字图书馆工作将起到推动作用。

3.6 生成方式:由手工向自动化方向发展

机读目录格式时代,大多数用于资源发掘的原

数据都是经过耗费人力的图书馆编目完成的,虽然图书馆编目至今仍然是查找书刊资源的最佳标准,但对网上资源来说,耗费太大,不可能在实践中实现。在另一个极端,检索引擎则用机器在短时间内处理大量搜索来的网页,做成索引文档,以供检索,从其成本—效益之比来看可以说走了一条十分成功的道路,但其质量可想而知。在这两个极端之间,是否有新的道路?经计算机处理的文件(如Word文件、PDF文件、网页、TIFF图像等)本身带有多项元数据,例如大多数编辑软件都支持用户自己输入元数据(标题、关键词、作者隶属单位等),能自动记录一些元数据(生产日期、有标记的文件结构、文件生成中采用的软件名、文件格式等),还有一些高级算法可以对文件按内容自动分类,或与内部分类词表匹配,自动给类目关键词。所有这些若能相结合,有不少元数据可以做到自动生成。采用XML为元数据格式、使用XML Schema描述元数据标准也促成了数据处理和转换的自动化。比如使用JAVA和Oracle数据库系统可以通过共享XML Schema,将XML格式的元数据直接装入数据库,或将数据库记录输出XML元数据文件。

自动抽取、生成元数据是当前重点研究对象,是未来趋势。前面所谈到的几个特点,模块化、简单化、语义化等都为自动生成元数据提供了先决条件。

3.7 对元数据互操作的前所未有的重视和要求

多种元数据的出现,带来了元数据之间的互相转换的需求。由于不再以一(一个机读目录格式)当十,众多元数据标准和元数据集、定义等共存于信息世界的有限空间。就好像过去大家曾被强迫学说同一种语言,现在一经“解放”,许多新的语种被创造出来,在不同规模的团体中得到应用,人们不再局限于用一套规定的词汇和语法来创造各种作用,他们充分发挥想象力,为自己需要的作品形式定义自己的词汇和语法。但是,语言之间若互不沟通,最后的交流便不能实现,再好的作品也无法得到最大面的欣赏。同时,动辄产生一种语言(含词汇和语法),也是很耗人力物力的。

为了排除这两方面的隐患,元数据之间互换性的工作得以在以下几种主要方式下进行:

(1)制作元数据之间的对照表(Crosswalks)。目前大多数比较通用或有名的元数据格式之间都有对照表,取一种元数据集为源格式(Source),将其他的目标(Target)格式中的元素一一与源格式相对照。

(2)在已有元数据集的基础上扩充,既保证了专用性,又保证了兼容性,而且省却许多重新定义的过程,比如在 DC 元素集上加入几个元素而形成用于教育的元素集(DC-ED),和在教育专用 DC-ED 上再进一步扩展的毕业论文元数据集。再如 DL ESE 在整个 IEEE-LOM/IMS 元数的九大类基础上,加入第 10 类,专门用于地球科学的专业元数据。

2002 年产生的 MODS,可以说是一套新的描述元数据格式,但它的产生完全是在 MARC 基础上的,首先,MODS 将机读目录格式中相关的字段、元素加以集中,仍如原来的 100 号和 700 号字段都与作者有关,现在被放在“作者”下。其次,MODS 不再采用数字符号,而采用与其他元数据相同的含语义的文字标记。MODS 的出现,将缓解前一段没有详细的描述元数据标准的矛盾。

(3)“资源描述框架”(Resource Description Framework,RDF)标准的产生。在 W3C 的标准中,有一项是 RDF,简单地说,RDF 允许在一条数据中“借”用不同元数据格式的元素,用这些分别选出的元素形成一个完整的格式。在每条数据前面,RDF 要求用 XMLNS(Name Space,XML 域名地点)指出这些元素的出处,通常是其元素集的元素定义的网上地址。从理论上讲,今后也许不需要产生新的元数据格式、元素集,因为人们所需要的元数据记录中可以容纳从现有的众多的元数据格式来的元素。

(4)METS。前面已经几次提到 METS,在这里再从互操作性方面来解释一下。METS 可以有 6 部分数据组成:

标示部(header):记录本身有关信息,如元数据制作者或单位,数据生成更新时间,数据识别符,以及数据状态。

描述元数据:METS 自己不定义描述性元数据,使用时可以引用任何描述性元数据标准如 DC 或 MARC。既可以是原始数据记录本身,也可以是一个记录的辨识号指向存贮在外部的(称为 external)元数据。

管理元数据:包括在这里的有技术性、信息来源、版权管理、或数字信息校对等类型的元数据。它们分别来自不同类型的元数据标准,例如数码图形式美国标准或国会版权元数据标准。如同描述元数据,记录本身可以是一个实际记录或是连接到其他

数据库的链接符。

文档部分:可以数字文件(TIFF 或 JPEG)直接与元数据存在一起,也可以将数字文件的地址存储于此。相关信息可以按文件格式排列。

结构部分:用来管理多个数字文件的相互关系,例如图型文件与音响文件的对应关系,扫描文件的页码顺序。

行为部分(Behavior Section):METS 提供了一个新的途径将播放数字文件所需要的软件信息与元数据存储在一起。

4 结束语

Caplan 在 2000 年美国国会图书馆召开的“新世纪的书目控制会议”上形象地把元数据的发展形容成“百花齐放的花园,布满了许多的交叉走道,在一条陡峭而布满石块的路的顶上”。事实上,现在“元数据”一词的应用已经走出数字图书馆的界限,在工业界,元数据标准化也成为公司数据库或知识库的首要一环。我们尚不能预见元数据的发展还会形成什么样的局面,什么是理想的目的地。本文通过以上分析,希望给大家留下这样一幅画面,让大家看到远远超出图书馆目录、数字图书馆、数据库的范围的元数据发展和应用范围,这种范围在理论上可以说是无限的。

参考文献

- 1 Caplan, Priscilla (2000). International metadata initiatives: lessons in bibliographic control. Paper presented at: Conference on Bibliographic Control in the New Millennium, Library of Congress, November, 2000.
http://lcweb.loc.gov/catdir/bibcontro/caplan_paper.html
- 2 Duval, Erik, Wayne Hodgins, Stuart Sutton, Stuart L. Weibel (2002). Metadata Principles and Practicalities. *D-Lib Magazine* 8(4).
<http://www.dlib.org/dlib/april02/weibel/04Weibel.html>

曾 蕾 美国肯特州立大学图书情报学院副教授,图书馆学博士。

张 甲 美国斯坦福大学图书馆工作。

张晓林 中国科学院文献情报中心副主任、教授、博士生导师。通讯地址:北京市北四环西路 33 号。邮编 100080。

(来稿时间:2003-03-12)