

XML: 图书馆的战略时机*

□ 迪克·米勒 著 李 樱, 张 甲 译

XML(可扩展标示语言)以它用于数据和文件交换的万能格式而迅速地获得各界的青睐。事实上,XML正在成为信息时代的通用语言。目前由于互联网的飞速发展,“图书馆信息”已处于很不利的状况。为什么?尽管图书馆已经开发了联网目录,进行了特定馆藏的电子扫描工作、数字数据的收藏,以及建立了丰富的网页,但是图书馆信息与网络信息仍然保留着“数字分离”状况。图书馆收藏的数据信息的核心是以一体化图书馆系统所有的形式以及以MARC格式存贮的。无论是图书馆系统形式,还是MARC格式,它们的使用主要局限于技术服务部和系统图书馆员们所在的部门。即便是这些图书馆员,他们也很难将这些丰富的信息提取并与馆藏以外的信息融合。图书馆信息的分离给我们造成很多困难,例:从MARC数据中生产出标准的书目数据,自动生成某一专题的新文献目录(包括新的网络信息资源),书商之间相互交换数据,以及从某一图书馆系统转到另一个系统。

为什么我们要用自设的枷锁将我们的潜力束缚而畏缩不前呢?很多图书馆安装了一体化系统,但他们很快就承认有些实际问题的解决方法就是自己动手。于是,我们聪明地预示了标准化的要求,然而随着互联网的降临,我们很快就发现“我们”的收藏与以新的数字形式出现的如洪水般涌现的信息资源形成了对立的两大阵营。如果我们不尽快将图书馆信息融于主流的网络信息,我们就会处于被忽视的危险之中。

XML的突起

XML的异乎寻常的随意性、通用性以及功能集中使它提供了一个前所未有的机会。商业界认识到用户愿意寻求或查找单一的信息资源,所以他们就全力以赴地把具有诱惑性的信息与他们的招牌信

息放在了一起。而我们图书馆的资料一向以公正、诚信、全面、具有长期收藏的价值以及缺少商业气味的良好声誉。我们不应犯有损于我们名声的错误。有些专业社团组织做了些尝试,但不幸的是并未获得好的效果。图书馆可以将XML的技术加入自己的“武器库”,而且远比商业界要做的容易得多。这个事实也可以用于说服商业公司的用户来分享我们高质量、高水平的信息资源。

开放的互联网标准

XML正在成为用来表述信息内容(不是格式)的互联网标准,这个标准可以最有利地传送网络信息。从技术上讲,XML是万维网联盟(W3C)于1988年2月创建的一组规范。它是一种元数据语言,用于定义不限定数量的特殊标识语言。每种特殊标识语言会包括不限定量的标签,因此是可扩展的。XML源自SGML(标准通用标识语言)。SGML于1986年由国际标准化组织(ISO)批准认可,广泛地用于欧洲和出版业以帮助文本文件出版物的电子传送。

到了1996年,显然是因为SGML的复杂程度使它难于用于网络的应用开发。HTML同样在数字信息表述上有很大的局限性,加上流行的浏览器的生产厂商为了商业用途加入很多不规范的标签,由此造成了使用这些标签的网页与浏览器互不兼容的现象。为了解决这类问题,XML将SGML简化,保留了其有效的功能,采用了HTML的通用文法,提高了网络的效力,并对XHTML(可扩展的HTML)的出现产生了影响。

简洁就是优雅

经过编码以后的XML相应的文件/记录具有逻辑和物理两种结构。从逻辑上讲,它们是由一组

*英文原注: Copyright 2000, used with permission of Library Journal, a publication of Cahners Business Information, a division of Reed Elsevier. 该文原载于2000年美国《图书馆杂志》网联夏季增刊:XML:Libraries' Strategic Opportunity, Dick R Miller, *Library Journal*, New York, Summer 2000, pg.18.有关详细信息,见 <http://www.ljdigital.com/netconnect.asp>.

有上下等级的,经过命名的元素组成。这些元素在类别上大致相同,其层层下属的元素又与其小类相从。一个文件的每个实体都有一个根元素,其他元素则隶属于这个根元素。包含具体文字或其他内容形式的元素称之为容器元素(Container)。这样的结构很象一棵倒长的树,一个树根,很多树枝,树叶则是代表实际上的数据值。如同 HTML 一样,XML 的元素一定要由一对三角括号“<>”来定界(起始标签/结束标签)。与 HTML 不同的是 XML 必须使用结束标签。一个文件只有当它的元素由起始和结束标签标识,而且上下层次适当时才能说它组成形式完好。不象 HTML,XML 必须组合严谨(见图 1)。

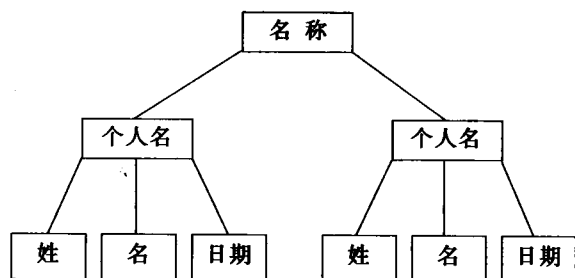


图 1

XML 的层次结构:XML 的元素最容易形容为一棵倒长的树,树叶是实际的数据值(在最底层)。

为了提供有关一个元素的性质,名称属性可以放在它的起始标签里。例如,一个唯一标识符可以被另一个属性用在一个文件中作为参引,这是标识符(ID)参引,见图 2。

```

<subject scheme="MeSh" type="topical" Level="primary">
  <descriptor id="12345">Liver Diseases</descriptor>
</subject>

<subject scheme="医学主题词表" type="专题" Level="主要">
  <descriptor id="12345">肝病</descriptor>
  <qualifier id="67890">药物治疗</qualifier>
</subject>
  
```

图 2 XML 部分样本

内容元素“subject”有三个属性(系统,类型,水平)和两个附属元素,每个元素附有一个标识(ID)属性。注意内容元素有其属性,但除了数值之外,它还包括其它承袭这些属性的元素。在 XML 中,元素名称和属性是很随意的,不过需要起始和结束标签。

从物理角度讲,实体可以允许一个文件的组成部分被分别命名和存贮。信息可以重复使用,参引非 XML 数据,例如图像。通常实体要在一个文件头部申明,这样才能在文件中被参引。其他 XML 的特性更神秘,不过已超越本文范围了。

太灵活了吗?

XML 的灵活性会导致混乱无序吗? 不一定。兴趣相同的组织可以开发一套 DTD (文件类型定义) 以适应共同的需要,同时扩展一些适用于本专业领域需要的文件类型定义。一个文件类型定义申明每一个经过许可的实体、元素、属性,以及它们之间的相互关系,这样就基本形成了一个用于相联的 XML 文件的逻辑结构的范本。它表述了数据的层次关系、个性、可准允的属性值,以及元素是否可选择、可重复等等。当一个 XML 文件配有一个 DTD 时,可以说这个文件就得到确认了。其实即便是它没有确认也已经很好地组成了。

DTD 并不是一定要有的,浏览器可以正确地识读不带 DTD 的文件,但是从完美的意义上讲,XML 文件还是应该有 DTD 相随。我们可以将 XML 编译软件定位,使其强迫 XML 文件与 DTD 相连在一起。

尽管分析数据和建立逻辑关系要求相当的知识,我们还是受到其他组织编制的 XML 实例的帮助,这些实例克服了习性和分歧而求得了共识。例如,DTD 已被用于定义生物序列标示语言,天文学、化学以及数学都有各自的标示语言。更有一种音乐标示语言支持声音以及以音乐文字符号为主的编码。

分工与合作

XML 实际是一组规范的核心。这组规范里的每一种规范适用于处理网上文件管理的某一具体方面。这些附属的标准正处于各自不同的发展阶段。例如,XSL (可扩展的文体格式语言)与级联式页面(Cascading Style Sheets)很相似,但两者在 XML 所包括的内容设计上又给予不同的显示指示。对内容采取不同的展示方法是 XML 强有力的一面。同一种数据可以根据使用的目的而定义不同的文体格式,从而以不同的方式显示。

其他语言规范包括 XLINK,XML 的链结语言。用于调节文件之间的相互链结。它已经超越了允许

从单一文件链接到多个相关文件的简单参见或互见。一个 XML 名称空位是一组以元素类型以及属性名出现在 XML 文件上的类名。这样在经过定义的上下文中,各个名称就保证是独一无二的。因为一个 DTD 定义一个名称空位,当从不同 DTD 而来的元素需要在同一个文件中出现时,名称的唯一性就产生作用了。

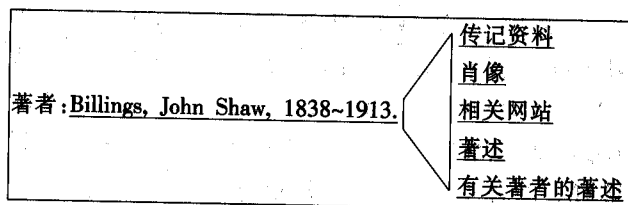


图 3 智能链接

这是一个目录或网站条目中的假设的部分,它显示了如何点击参见链接就可以给寻找有关这个著名的医生/图书馆员的读者提供多种有关资料信息的选择。XML 允许内含信息查找器在文献中建立所需功能,如有条件链接到范围注释,权威记录中的相互参见以及引发的数据库检索。

为高效率奠定基础

XML 具有有效、精确、可控和灵活性,由此对图书馆员产生了强烈的吸引力。同时 XML 的可伸展性(适用于新的标准)、复杂精确的链接、可调节性(允许在需求不同的多种上下文中重复使用信息),和它的相对简易使用为它未来的发展提供了统一的基础。XML 的标签直观易懂,非专业人员也可使用。

此外,XML 适用各计算机系统平台及应用软件,这就使它成为处理各种文件记录的理想工具。它既适用于结构严谨的数据记录,也可以用来处理不太规范的文件。XML 开始主要应用于网络出版业,但后来由于它对交互操作的支持而被商业界和科学等领域广泛地接受。对图书馆来讲,XML 正在取代 EDI(电子数据交换)标准。不同的数据资源在共用 XML 语法时,其数据更容易相融和得到处理。

XML 支持的新产品几乎是日有所见。2000 年 5 月,ORACLE 发行了人们期待已久的 IFS(互联网文件系统)。IFS 可以据 XML 的 DTD 自动地建立起一个关系数据库,并且可以很容易地输出 XML 文件。同样在五月,INTEL 宣布了他们的新产品可以在硬件层面上加快处理 XML 文件的速度。Sun 公司也为

Java 增加了延伸部分以适应 XML。这个延伸部分并入 SAX(用于 XML 基本 API),成为一个通用的 Java-XML 界面。微软公司的互联网探索者第五版是第一个支持 XML 显示的浏览器。Netscape 也宣布要将它的浏览器转成一个 XML 平台。除了业界的热门讨论外,很多公司都在积极地行动。

很多开放源码软件都可免费获取,包括 SAX 和 W3C 正在研制的 DOM(文件模式),一个与 XML 数据相接的标准界面。DOM 将会允许程序以及指令程序使用和更新文件内容、结构和格式。文件从而得到进一步处理,处理后的结果还可以被并回原来的网页。我们都知道会有越来越多的产品出现,使建立、使用和处理 XML 数据更为容易方便。还有一个相关驱动力,这就是经济。有人估计使用 XML 会将数据发行的费用节省下 30%至 60%。以下的例子可以说明 XML 广泛应用于图书馆的趋势以及潜在的效益。

电子文献保护

不同于 SGML,XML 使用固定的字符集 Unicode。Unicode 包括传统的 ASCII 代码字符,并采用双字节的方法使编码范围扩大到 65000 字符数。XML 现采用国际标准 ISO 10646 编码方案,可容纳多达二亿多个独立字符。这个国际标准允许 XML 使用计算机操作系统和应用软件在普通文献中处理注音符号、特殊字符,以及非罗马字符。这对图书馆界或数据网络的国际化尤其重要。

Unicode 的支持和不受计算机硬件平台的限制,XML 能够充分保证数据长期保存而不受未来硬件、软件以及网络协议的变化影响。斯坦福大学雷恩医学图书馆正在将电子邮件以及其他的一些由淘汰的软件生成的电子文献转成 XML 文件。雷恩图书馆有一批保存在磁带上的电子资料,它们是从 70 年代以来从事人工智能专家系统研究的学者,诺贝尔奖得主 Joshua Lederberg 以及 Edward Feigenbaum 所写的电子邮件、研究项目和计算机程序。由于这批磁带正在老化,需要对其进行转换保护。通过对电子邮件的分析,把它归入各类元素,例如日期、送至、收到、回信和信文等。接信人和送信人的姓名可以前后移置,相关信件的送收关系也可以用 XML 的链接功能加以保持。可用不同计算机操作系统的 XML 电子邮件系统将会免除这种信件

转换的需要。

书目数据库

数据库管理是个很复杂的话题,XML对复杂的数据结构提供了清楚的定义。名称空位可以用在DTD的联合部分以帮助理顺那些复杂的关系。支持XML的数据库界面已经在市场出现,如Oracle8i和Ozone。

XML用于数据库管理可行性的最具有说服力的例子就是美国国家医学图书馆。作为现代化医学图书馆计算机系统工程的一个部分,美国国家医学图书馆选了XML作为MEDLINE书目引文数据的格式,并在今年内将其1100多万条记录转换成XML的格式。该项计划的目的是将XML格式的数据作为关系数据库的输入和输出的数据格式。

美国国家医学图书馆利用数据格式转变的机会,对数据进行了重组。例如分隔错误记录,从篇名提取信息,提供新的元素如团体著者等。2001年起,XML将成为美国医学数据库的唯一发行格式。美国国家医学图书馆已经计划出版医学主题词表的XML版,它的MARC记录最终也会采用XML。XML还被用于美国国家医学图书馆新的网关,一个能够查找美国国家医学图书馆多种终端检索系统的智能查寻工具。

美国航天太空管理局的天文物理数据库也选择XML重新格式化其书目记录。DialogWeb已开始使用XML数据库界面,世界知识产权组织则宣布XML是他们接受专利申请件的首选格式。

XML可以对检索结果进行再处理,索引的输入项能够在动态中从具有结构和功能的XML格式的检索结果中自动生成。例如在著者姓名显示时,可以加一个“打开”的选择项供读者查找与该著者相关的链结和有关这个著者的记录数目。

尽管上下层次结构简单,XML却展现了它与复杂的书目数据的显而易见的可融性。鉴于XML的很多优点,图书馆员必须对MARC和AACR2作个认真的审视。为MARC写出DTD是个相当大的工作,但是我们怎能对一个为六十年代卡片目录设计的格式而无动于衷呢?将MARC的长处并入一个现代的、和编目规则相关联的二十一世纪的现代格式

是安全的和可能的。

雷恩图书馆在1998年9月初开始了一个Medlane项目,将书目记录转为XML格式并与其他网上资源合为一体。雷恩馆写出了DTD试本并重新结构和简化了MARC,同时还发行了XMLMARC软件,示范了MARC转换为XML格式的可行性。目前,Medlane项目集中在索引及用户界面的开发上。

法国一个政府机构发行了BiblioML,将Unimarc转换成XML。国会图书馆也出版了一个从1995到1998年的MARC与SGML的文字对照。Logos研究系统公司的从MARC到XML,XML到MARC的转换与对照工具也已见诸文字。

一个完美的解决方案

在三到五年内出现一个完整建立在XML基础上的图书馆一体化系统是有可能的。图书馆系统开发商们,尤其是Endeavor,正在开始将XML放入现存的系统中。美国图书馆学会可以用W3C的模式加速发展所需要的标准,特别是DTD或图书馆系统数据规则的建立,包括:书目、规范文档、馆藏、读者、书商、借阅记录、馆际互借、流通记录等。财务方面的DTD已经有了。尽管这是一个艰巨的任务,但在其他领域类似的工作已开始,为什么图书馆系统就不可能呢?

XML图书馆一体系统

XML初看起来是复杂的,但对不易解决的问题提供了一个完美的解决方案。如果图书馆一体化系统和MARC能以XML为基础,我们就可以站稳脚跟,集中精力提高系统的功效和系统的兼容性,使之远远超越今天的系统。这是一个一次性的改变,朝着开放通用格式的改变。这个要有系统厂商参与的改变可以使受时代限制的MARC格式重新焕发青春,同时将图书馆员置于能更好地为读者服务的主流位置上。

著译者单位:美国斯坦福大学雷恩医学图书馆(Lane Medical Library, Stanford University Medical Center)

收稿日期:2001年5月18日