

XML在数字图书馆资源共建共享中的应用

内蒙古自治区图书馆 周虹利

〔摘要〕 本文介绍了XML的概念、特性,根据XML的特性和优势对XML在数字图书馆资源共建共享中的应用进行了阐述。

〔关键词〕 XML 数字图书馆 资源共建共享

数字图书馆资源目前以分布的形式存在于各种机构的数据库或者是文件服务器中,由于各自的系统平台、数据库、数据结构、类型定义、资源描述以及资源的最终形态都有很大的差异,因此,要实现资源的共享就必须进行资源描述的标准化。作为第二代Web语言,XML以其良好的语义和清晰的结构受到人们推崇,已经成为网络间进行数据交换的理想标准。只要按照一定的规范用XML描述各种数字图书馆信息资源,就可以很好地实现数字图书馆资源的标准化,从而在一定程度上解决数字图书馆资源共享问题。

一、XML的概念

近代图书馆理论与实践是奠基于图书馆自动化的发展,而图书馆自动化则主要又随着信息技术和计算机技术而改变,因此图书馆界一直存在着“不遗余力地追逐信息技术和计算机技术”的特性。图书馆自动化中一个核心的问题是“数据交换”,而XML对网络数据交换提出了一个新的解决方案,XML(Extensible Markup Language),中文译名为“可扩展标识语言”,由W3C(万维网协会,the WorldWide Web Consortium)于1998年2月正式推出XML1.0版本。XML不是

一种编程语言,它和SGML一样是一种元语言(meta language),它不像SGML那样的复杂,又比HTML有严格的语法规则,可以说集中了二者之长,却无二者之短,预见将会取代HTML的地位,成为网络描述语言的主流。XML是跨平台、跨网络、跨程序语言的一种数据描述语言,在数字图书馆共建共享中具有很高的发展潜力。

二、XML的特性

1. XML可以实现数据形式和内容的分离

XML将各种结构化,半结构化和非结构化数据集成起来,在XML平台上整合应用,使杂乱无章的信息海洋得到根本改善,每个数据节点将实现信息有序存储,并能为对方接受,使数字图书馆的智能搜索引擎,数据挖掘与数据分析工具能够得到很好的实现,真正实现用户的个性化服务。XML将成为数字图书馆最重要的基础性语言。

2. XML具有数据交换的特性

用HTML制作的界面要和数据库打交道时,都要经过复杂的转换手续,而数据与在数据库与数据库之间交换时,更是要大费周折。由于XML是结构化的数据,所以要储存或是要在数据库和数据库之间交换时间,

都非常容易。XML 今后将会是数据在数据库和数据库之间交换时的标准方式。

3. 整合不同数据库

由于 XML 具备了数据库字段的作用,因此 XML 也相当适合在网络上整合不同数据库,以往不同数据库的内容要能够互通,必须采用相同的格式,或者各自转成相通的中介格式,才能达到互通的目的,如今大家就可以使用 XML 当作中介格式,来达到不同数据库互通的目标。

4. XML 解决显示格式问题

XML 定义了新的标识语言,能够应用到例如音乐乐谱、化学方程式、数学公式、财务电子表格,以及工程应用等各种不同的专门领域,再通过标准的客户端处理程序,就犹如现行的 Plugings 程序一样,就可以达到在不同的浏览器或不同的 HTML 规范版本下相同显示的目的,让每位浏览者都能看到相同的文件内容;也可以将一段相同的文字内容,以各种不同的样式呈现,或者动态地改变文件的内容(就如同是 IE 中强调的动态内容),使得 Web 上的文件能够具备多样化的表现。

三、XML 在数字图书馆资源共建与共享中的应用

从 XML 具有的特性可以看出,XML 能够解决数字图书馆资源共建与共享中的海量信息资源的加工、存储、交换等问题,也能够满足实现数字图书馆建设的不要求。目前 XML 在数字图书馆资源共建共享主要具体表现在以下几个方面:

1. XML 可实现数字图书馆 Web 信息资源整合

XML 应用于 Web 的最大长处是它与 DOM(对象文件模型)的接口。数字图书馆 Web 信息资源整合的技术思路,就在于建立统一的数据交换标准和接口,以保证异种库之间的透明访问。目前,在国内高校应用广泛的 Web of Science,就是对于网络数据标准

接口模型的很好诠释。对于 Web 信息资源整合来说,XML 技术具有诸多优点:其一,XML 允许组织、个人建立适合自己需要的资源集合,可广泛应用于信息交换的多种领域;其二,XML 把文档的三要素独立开来,其自我描述性质能够很好地表现许多复杂的数据关系,使得应用程序可以在 XML 文件中准确高效地搜索相关的数据内容;其三,在信息发布方面,同样的 XML 接口,可以适用于不同的用户端访问形式;其四,XML 独立于平台,有利于跨平台间的信息交流,完全可以充当网际语言,并有希望成为数据和文档交换的标准机制;其五,XML 能够更准确地表达信息的真实内容,其严格的语法降低了应用程序的负担,也使智能工具的开发更为便捷。

2. 元数据 DC 的著录与存储

目前,数字图书馆的主要编目模式是 DC(Dublin core,都柏林核心元数据)编目模式。DC 编目模式是为描述网络资源、支持网络检索而建立的元数据格式,现已成为 InternetRFC2413 和美国国家信息标准 Z39.85。目前,国际上主流的数字图书馆方案基本上都采用了这一格式。普遍作法是:采用都柏林核心集作为元数据的语义规范、RDF(Resource Description Framework,资源描述框架)作为语法规范,以 XML 为表现或存储形式,将数字资源存储在计算机系统中。

例如,下面是利用都柏林核心元素集定义一个图书馆目录的 XML 简单实例:

```
< ? xml version = "1. 0" encoding =  
"UTF - 8" >  
< ! Doc TYPE Bibliographic biblio. dtd  
>  
< Bibliography >  
< HEAD >  
< TITLE > Dublin Core 形式节目 < /  
TITLE >  
< / HEAD >
```

< BODY >

< dc : Title > 全国专业技术人员继续
教育公需科目教材 < / dc : Title >

< dc : Creator > “edt (主编)” > 白春
礼

< / dc : Creator >

< dc : Publishe > 中国人事出版社

< / dc : Publisher >

< dc : Subject > 创新能力建设 < / dc
: Subject >

< dc : Description > 创新是一个民族
进步的灵魂, 是一个国家兴旺发达的不竭动
力。提高我国自主创新能力, 建设创新型国
家, 是国家发展战略的核心和提高综合国力
的关键。 < / dc : Description >

< dc : Contributor > 中国人事出版社

< / dc : Contributor >

< dc : Date > 2009 - 6 - 1 < / dc :
Date >

< dc : Type > 专业技术人员创新案例

< dc : Type >

< dc : Format > Ebook < / dc : Format
>

< dc : Identifier id = “xyz” scheme =
“ISBN” > 978 - 7 - 80189 - 851 - 7 < / dc :
Identifier >

< dc : Source > = “http : / / www. zjg-
su. edu. cn” < / dc : Source >

< dc : Language > chi < / dc : Language
>

< dc : Relation > 内蒙古图书馆 < / dc
: Relation >

< dc : Coverage > 专业技术人员创新案
例

< / dc : Coverage >

< dc : Rights > 中国人事出版社 < / dc
: Rights >

< / BODY >

< / Bibliography >

从上面这个例子可以看出, XML 的这种
描述方法不但适用于机读目录, 并且对数字
图书馆建设人员来说也十分简明易懂, 还为
都柏林核心元数据的发展提供了技术上的
支持。

3. 资源描述的标准化

现代数字图书馆未能完全实现资源共
建共享, 分析其原因, 主要有两个方面: 一是
没有统一的资源描述标准; 二是没有统一的
资源描述方式。对于这两者, 正是 XML 最大
的优势所在。在数字图书馆领域, 我们可充
分利用 XML 可扩展性、灵活性、自描述性等
优点, 很好地解决数字图书馆系统间的集
成、资源发现、系统性能的瓶颈等问题。构
建基于 XML 的数字图书馆体系已经成为数
字图书馆的一个重要发展趋势。由于 XML
可用来描述信息及对其进行组织, 所以可以
将它当作一种数据描述语言, 用它来描述数
据成分、记录和描述数据结构, 甚至复杂的
数据结构。可以用 XML 方便地创建出共享
的自定义数据结构, 生成有关服务、产品、商
业交易以及网络教育的结构化信息。这些
信息是可以在网上进行交换的。也就是说,
用 XML 能描述一个过程, 原封不动地移动数
据, 重新对信息进行打包, 让这些信息更适
合特定的信息接收者。如此一来, 只要按照
一定的规范用 XML 描述各种数字图书馆信
息, 就可以实现数字图书馆信息数据结构的
标准化。

4. 在 MARC 中的应用

除 DC 编目模式外, 另一种是机读目录
编目格式。机读目录 (machine readable cata-
logue 简称 MARC) 是发展历史比较长的元数
据格式, 是为描述、储存、交换、处理及检索
信息资源而精密设计的基础。MARC 格式是图
书馆信息资源编目的基础。MARC 格式是图
书馆中的馆藏资源的主要表示格式, 它提供
了一整套完整、详尽、复杂的流式数据表示
规范。但是由于 MARC 元数据格式的专用性,
读者

必须依靠专用的客户端和图书馆系统所提供的检索工具进行资源搜索, MARC 格式却成为图书馆数据资源的整合进入网络流通构成的最大障碍, 目前, MARC 在句法层面进行了 XML 化改革。其具体做法是: 应用 XMLDTD 机制或 XMLSchema 机制定义 MARC DTD 或 XMLSchema 来解决 MARC 的类型字段及字段标识。在 XML 中, 除了一般的语法限定外, 最重要的也是 XML 用户扩展的重要途径之一, 便是 DTD (Document Type Definition, 文档类型定义)。XML 的 DTD 机制就是为了定义逻辑结构的限制和支持预定义存储单元的使用。定义 MARCDTD 可以解决 MARC 类型标识、字段标识和子字段标识的问题, 从而让 MARC 数据从严格复杂的规范流格式数据转换成机器可读的 XML 结构化数据, 将 MARC 数据转化为 XML 文档, 实现 MARC 书目数据库和 Internet 上的非书目数据库的集成成为可能, 从而使得现有的大量的 MARC 格式书目数据能方便地在数字图书馆中加以利用, 从而实现 MARC 数据的 XML 结构化, 进而实现 MARC 数据库与其他 Web 信息的集成, 这在当前数字图书馆建设中具有重要意义。基于 XML 的 MARC 是一个面向 WWW 的开放式书目数据格式。它的出现是图书馆的重要发展机

遇, 使图书馆系统在网络环境下呈现出勃勃生机, 在图书馆业务工作中实现了真正意义上的网络采购和各种格式的书目资源转换。

四、结语

由于当前数字图书馆信息资源呈海量级发展, 但存在着针对性不强、重复建设等现象, 造成了资源的严重浪费。解决这个问题关键就在于信息的标准化, 而 XML 和元数据这一新兴的技术, 恰恰能有效地描述信息资源, 实现资源发现和交流, 很好地解决数字图书馆资源共建与共享的问题。从发展趋势来看, XML 和元数据作为一种数据交换形式在 Internet 上已得到广泛使用, XML、元数据也必将在数字图书馆中得到应用。利用 UML 和 XMLSchema 创建数字图书馆元数据的方法, 通过元数据把复杂的信息方便地组织起来供用户使用, 可极大提高数字图书馆系统的效率, 进一步促进数字图书馆资源的共建与共享。

参考文献:

1. <http://www.w3.org/XML>
2. 高峰/MARC 数据转换为 XML 文档的设计与实现/现代图书情报术/2005, (1): 22-25
3. 景民昌、王平/基于 XML 的数字图书馆 WEB 应用开发/计算机与化/2004, (6): 76-78

(上接 41 页) 提供馆内文献资源, 开展网上借书、网上咨询服务, 使得读者足不出户就可以知道图书馆里是否有自己想要的书刊信息。

随着我国公共文化服务体系建设的推进, 图书馆不能满足于开展阵地服务、传统服务, 而要充分利用馆藏文献和设施等条件, 利用现代化信息技术, 积极扩展图书馆社会教育功能, 扩大服务覆盖面, 为社会公众提供多样化、个性化服务, 提高图书馆的文化服务能力。

参考文献:

1. 老世龙/经济欠发达地区公共图书馆延伸服务模式研究/江西图书馆学刊/2010(2)
2. 杨华芳, 龙小玲/县级图书馆在农家书屋建设中的作用探讨 - 以江西省吉安市为例/江西图书馆学刊/2010(2)
3. 傅斌/依托文化共享工程服务农村留守儿童 - 公共图书馆为农村留守儿童服务的思考/江西图书馆学刊/2010(2)
4. 李姝/对图书馆延伸服务的有益探讨/河南图书馆学刊/2010(2)