

击 Order Documents 键;点击 Related Articles 键,显示相关文献;点击 Expand 键,以扩展格式显示该条文献。扩展格式页上的 Publisher Link 键,用于链接到文献出版者的网站。

Download or Display 键用于多条文献的显示、下载和通过 E-mail 发送文献。点击该键,进入下载或显示页。“Fetch”下拉菜单用于选择每页要显示的文献条数;“Results from”下拉菜单用于选择数据库、信息系统或 Locker;“Details”下拉菜单用于选择文献记录详细程度(包括简单、扩展和完全);“Format”下拉菜单用于选择检索结果的显示格式(包括无字段名称、有字段名称和有字段标识的输出格式);“Destination”下拉菜单用于选择显示或下载的目的,如果为了打印,应选择“Display for Printing”,待显示后,点击网络浏览器工具栏的“打印”即可;“Download or send as”下拉菜单用于选择下载的

文件类型;如果想把检索结果发送到电子信箱里,首先要在“Destination”下拉菜单中选择“Send Via Email”,然后在该页左下方的地址框中输入 E-mail 地址,两个地址之间用逗号隔开。此外,如果想给通过 E-mail 发送的检索结果加些说明,可在该页右下方的空栏中输入。各项选择完成后,点击 Go 键即可。

参考文献:

- [1] Help. [http://gateway.nlm.nih.gov/gw/Cmd?GBMasicSearch\[DB/OL\]](http://gateway.nlm.nih.gov/gw/Cmd?GBMasicSearch[DB/OL]). 2002-05-02.
- [2] Overview. [http://gateway.nlm.nih.gov/gw/Cmd?GBMasicSearch\[DB/OL\]](http://gateway.nlm.nih.gov/gw/Cmd?GBMasicSearch[DB/OL]). 2002-05-10.

[本文编辑:杜海洲]

XML 在数字图书馆中的应用

申飞驹,袁红,董建成

(江苏省南通医学院图书馆,江苏南通 226001)

摘要:介绍 XML 的特点和功能,并从有序地组织信息资源、高效地满足用户需求、有效地保护知识产权、更好地共享信息资源四个方面分析了 XML 在数字图书馆中的应用,及在国内数字图书馆中的应用现状与前景。

关键词:SGML;HTML;XML;数字图书馆;网络环境;信息资源

中图分类号:G250.76

文献标识码:A

文章编号:1671-3982(2002)06-0040-03

Application of XML in digital libraries

SHEN Fei-ju, YUAN Hong, DONG Jian-cheng (Library of Nantong Medical College in Jiangsu Province, Nantong 226001, China)

Abstract: The characteristics and function of extensible markup language (XML) are introduced. Its application in digital libraries is discussed from 4 aspects such as arranging information resources in order, satisfying user needs efficiently, protecting intellectual property rights (IPR) effectively and sharing information resource. At the meantime, the current status and prospects of its application in China are outlined and forecasted.

Key words: Extensible markup language; Digital libraries; Networked environment; Information resources

数字图书馆是一项重要的国家信息基础设施,将高新技术运用到网络信息资源的生产、交换和使用中是数字图书馆的一项基础工作。高新技术要解决的问题是面向 21 世纪的因特网,有序地组织信息资源,高效地满足用户需求,有效地保护知识产权,更好地共享全球的知识资源,切实解决目前在因特网上长时间寻找、太多的垃圾、不易实现资源共享、难以保护知识产权等问题。

XML(Extensible markup language)是 W3C(the World Wide Web Consortium)工作组制定的可扩展标记语言。作为未来网

络的基础技术,XML 在数字图书馆中的应用已不再是纯理论层次上的概念,而是不断深入和发展中的现实。本文分析 XML 的特点和功能以及在数字图书馆中的应用、现状和前景。

1 XML 的特点和功能

标记语言(HTML, SGML, XML 等)使用标签及其他标识来描述某个事物应该如何合理地显示在计算机屏幕上。得益于其简洁和通用的特性,HTML 是目前使用最广的标记语言,但它仅描述信息表象而不说明信息本质,且它的标签集合是固定的,其致命弱点在于“仅为满足发布”。

SGML(通用标记语言)作为 HTML 和 XML 技术的前身,是一种独立于各平台的语言,用于规定信息的语义结构。但由于 SGML 过于复杂,Web 浏览器解释 SGML 语法规则的工

收稿日期:2002-02-28

作者简介:申飞驹(1966-),男,江苏南通人,本科,馆员,参编著作 2 部,发表论文 10 余篇。

作量过大,限制了它在网络中的广泛运用。

XML 是 SGML 的一个经过严格筛选的子集,它以 SGML 20% 的难易度实现了其 80% 的功能。同时,XML 解决了 HTML 的许多限制和不足,用户可以自行设立专用标签,如在医院的诊疗记录文件中设立 < 诊疗记录 >, < 病人 >, < 诊疗日期 >, < 治疗方案 > 等标签描述内容元素,并通过标签的层次结构进一步揭示各内容元素之间的关系。所以,XML 被称为元语言(Meta Language)的正式标记语言,它将处理的触角深入到了信息的内容层次。XML 文件通常有两部分:一是 XML 标签及其内容,二是定义标签及其相互关系的文件定义类型(Document Type Definition, DTD)。

如何进行信息资源的开发利用和共享已成为全球性的热点问题,人们期待研究和开发基于信息内容的新理论、新方法和新技术来解决开发、利用、共享信息内容的现实问题。XML 以其强大功能得到了广泛的认同和应用。凡有信息利用的领域,就可应用 XML。而应用 XML 的领域,如数字图书馆,电子商务等倍受青睐。

2 XML 在数字图书馆中的应用

早在 1995 年,国家图书馆就开始跟踪发达国家数字图书馆的研究,并先后在文化部和国家计委申报若干有关的科研项目进行研究。1999 年又建立了数字图书馆的实验环境,进一步验证建设数字图书馆的关键技术,包括数据资源建设的技术支持,采用标记语言对资源进行标记;元数据查询系统;形态数字资源的对象数据存储和检索系统;高速馆域网的网络系统;多家合作建设资源库的试验及实践。作为一项新的技术,XML 提供了制订数字图书馆技术方案的新思路。作为信息处理的新国际标准,XML 在数字图书馆的信息存储和加工、流通与交换、利用与再利用、安全与保护、扩展与增值等方面的应用都是其他标准所无法比拟的。从宏观上说,XML 在数字图书馆中的应用包括外部应用和内部应用。外部应用是指 XML 提供了对 Web 结构化信息管理的可能,使得数字图书馆参与网络信息资源组织与管理的环境更为有利;内部应用是指 XML 优化了数字图书馆信息资源的组织、使用、保护和共享。XML 在数字图书馆中的应用主要包括以下四个方面。

2.1 有序地组织信息资源 元数据是网络资源组织发展的热点,是描述和管理对象数据的数据。MARC 和都柏林核心集都是元数据。对于日益增长的网络信息资源来说,元数据具有重要意义。一是发现工具,它可提供有关资源本质和内容的信息;二是定位工具,它可提供网络资源的确定位置,极大地方便了网络资源的组织和资源发现。

数字图书馆的资源内容是按照用户需求挑选出来的高价值的知识信息,经过分类、编辑、整理、加工,以受教育者易于接受的形式提供。它是一种增值的海量信息资源库。为

实现海量信息的有序组织,数字图书馆将信息资源分为元数据和对象数据。元数据是关于信息资源的形式(格式)、主要内容、存放位置等信息的集合和摘要,相当于传统图书馆的目录。对象数据是指数字化的文本、图片、声音、影像等数据,相当于传统图书馆的藏书(内容)。充分利用元数据对信息的特征进行描述来构造某些信息的特征数据,利用“特征数据”高效有序地组织信息资源。

元数据与 XML 发展密不可分。元数据是为解决网络信息资源有序组织的效率与质量之间的矛盾而产生的。HTML 在描述非结构化数据上缺乏足够的功能。考虑到数字图书馆信息资源增长的速度,任何手工和人力著录信息资源的做法都是不可行的。元数据格式也只是权宜之计,因为任何一种元数据格式都无法对所有信息资源进行著录。但如果使用有语义的标识标记数据或文档,就可以使数字图书馆信息著录和组织走出困境,XML 正好适应了这样的需要。利用 XML 可以生成各种用于不同目的的元数据。由 W3C 提出的 RDF(Resource Description Framework,即资源描述框架)是一种基于 XML 的元数据规范,是在 XML 上更深层次的应用,采用 XML 句法表达语义。RDF 使数字图书馆具有更佳的搜索引擎功能,在数字图书馆的网络导航中将发挥巨大作用。

2.2 高效地满足用户需求

2.2.1 丰富信息表现形式。HTML 的语法格式过于简单,无法处理如音乐乐谱、化学方程式、数学公式以及工程应用等数据。就显示而言,不同的浏览器或不同的 HTML 规范版本,显示结果不尽相同,而 HTML 本身又无法扩充来支持各种专门的格式,因而大大限制了 HTML 信息表现的领域。XML 定义新的标识语言应用到不同的专门领域,再通过标准的客户端处理程序,就如同现行的插件程序一样,可以达到相同显示的目的,让每位浏览者都能看到相同的信息,也可以将相同的信息以各种不同的样式呈现,如 HTML 和 PDF 等,或者动态地改变信息内容。XML 给数字图书馆带来了丰富的信息表现形式,以满足用户多样化表现信息的需求。

2.2.2 方便信息获取。XML 应用在数字图书馆便于信息获取的机制在于查询系统直接针对元数据库。目前因特网上提供的信息资源标引深度多以文献条目为主,大多数全文文献则是不经标引上网,搜索引擎按每种文献的前 50~200 字进行关键词抽取,与查询请求匹配得出结果,既难查全也难查准。数字图书馆中信息资源的标引则深入到文献内容,对其每一数据对象使用 XML 进行标记,并抽取元数据组织元数据库。查询系统在元数据库中寻找满足用户要求的信息。例如,当今世界上每年有 750 万份文献,假定每份文献有 200 页,每页有 400 个汉字,需要的存储空间为 1 200G,如果在这 1 200G 的全文中检索关键词,所需的时间、检出信息的相关度等都是不可想象的。假如采用元数据方式,如果描

述每份文献需要 50 个字段,全部字段的长度平均为 1 500 字节,那么元数据的存储空间为 10.25G,在这 10.25G 的元数据库中查询文献的特征,使用的时间和检索出信息的准确率与简单的全文检索相比将极大地提高。如果将全国的元数据在一个大的区域中共享检索,可以有效地解决全国资源利用的关键问题,即让用户知道谁有什么对象数据。可见,在数字图书馆信息资源加工软件系统中采用 XML 技术,是实现数字图书馆高效信息获取的重要措施。

2.2.3 方便信息利用。HTML 只是将信息传达给读者,用 HTML 制作的界面与各数据库打交道时,要经过复杂的转换,而数据在数据库之间交换时,更要大费周折。而 XML 数据是结构化的数据,可通过程序进行各种操作,且 XML 平台上可使用多种软件,这就使得用它表示的数据具有良好的使用效率和可重用性。读者下载的 XML 文件可利用不同的程序自动加工,如正文某一部分的抽出和显示,存储或转存,在数据库之间的数据交换等。对于一篇文学作品来说,XML 在信息利用方面的优势可能并不突出,但对于文件中成组、规律性的统计资料、数据库内容来说,XML 能够将这些信息有效地利用。

2.3 有效地保护知识产权 知识产权的特点之一是“专有性”。而网络上的信息则多是公开、公用的,极难被权利人控制。数字图书馆的信息资源利用过程就存在着这样的问题,它突出地表现为对已建立的数据库的法律保护,为其使用设立合理的法定许可范围。

数字图书馆兼有公益服务和商业服务两方面的职能,不同的数据库或同一数据库的每一栏目,都有不同的读者群,针对不同使用者应有不同的使用规定。例如针对学校、政府、商业机构的不同性质,收费与否和收费多少可以不一样,使用的权利范围也有所不同,数字图书馆有必要分门别类地作出有区别的授权使用规定。数字图书馆信息资源,应从技术上保证不会被无权限者利用以及有权限者无限地传播,如自由下载,打印等。

数字图书馆知识产权的保护对象是对象数据库的集合,利用 XML 可划分知识产权保护的级别或权重,可注明 Web 页的知识产权要素,如权利,出版者,创造者,贡献者等;建立以各地的特色馆藏为中心的对象数据系统,并能有效地根据调度系统与知识产权、用户协议等因素将对象数据发向所需的用户。这些将对数字图书馆知识产权的保护起到积极作用。

2.4 更好地共享信息资源 XML 具有优越的共享性。XML 中的数据内容与具体应用无关,并提供不同类型信息资源之间的标准转换格式,从而可以方便地传送自我描述的数据和获取他人不同结构的数据。这一点使得人们利用数字图书馆可以更好地整合和共享不同的网络资源,在使用者和数字

图书馆之间维持一种更密切、更有效的关系。

3 XML 在国内数字图书馆中的应用现状及前景

3.1 我国数字图书馆的研发情况 国际上数字图书馆及相关概念出现后,立即引起国内有关部门的重视,并开展了相应的研究与开发工作,包括数字式图书馆试验项目,基于特征的多媒体信息检索系统的研究开发,SGML 的图书馆应用;中国试验型数字图书馆,知识网络——数字图书馆系统工程,辽宁省图书馆的数字化图书馆,教育部的数字化图书馆攻关计划等。

3.2 中国数字图书馆工程项目 国家图书馆于 1998 年 7 月提出该项目。它采用与国际同类主流技术有接轨前景的技术方案,如标准通用标记语言(SGML),统一资源名称(URN)等。

3.3 中国数字图书馆实验演示系统 为探讨建设中国数字图书馆的基本途径和方法,为中国数字图书馆工程开展前期实验研究并得到实践经验,国家图书馆于 1998 年底开发了试验环境及演示系统。该系统在多媒体数字资源的创意、设计和加工,元数据抽取和元数据库的生成,对象数据库建设,调度系统的选用,以及实现跨库检索和用户界面等方面进行探讨,取得了初步的实践经验。国家图书馆开发了数据加工、标记、管理方面的软件,开发了元数据和对象数据装入和管理方面的软件,开发了包括多媒体对象在内的动态页面生成软件,开发了支持 SGML 和分布式数据库检索、查询软件以及元数据和对象数据连接的接口软件等。

由此可见,XML 在国内数字图书馆中的应用还处于试验阶段,SGML 仍是数字图书馆的基础技术。实际上,XML 诞生之前用 SGML 生产和加工数字图书馆内容和元数据在国内外十分普及。但是,随着 XML 的优势的显现和相关工具软件的大量开发,用 XML 来记述数字图书馆的内容和元数据,必将成为数字图书馆技术发展的主流。我们欣喜地看到,基于 XML 为基础的 RDF 描述体系也实现了开放、标准、统一的内容管理,使数字图书馆的资源建设能在相当长的时间内得到保护,也为未来向新的体系与标准升级提供了便利。

参考文献:

- [1] 石英弘,李颖.未来网络的基磐技术——XML 的理论与应用[M].北京:华艺出版社,2002.1.
- [2] 罗良道.XML 及其在图书馆的应用[J].图书馆杂志,2001,(2):26-30.
- [3] 张咏.XML 及其在图书馆和情报检索中的应用[J].现代图书情报技术,2001,(2):30-34.
- [4] 赵英,党跃武.组织 Web 资源的新的标识语言——XML[J].图书馆杂志,2000,(9):22-25.

[本文编辑:陈 界]