基于 XML 的数字图书馆技术体系结构研究

The Technical Infrastructure of Digit Library Based on XML

郭景峰　陈玲灵

（燕山大学信息工程学院　秦皇岛066004）

Abstract　This paper presents a kind of technical infrastructure about digit library based on XML concept , which contains resource-dealing system, data management system ,user interface, data protect system.

Keywords　Digit library , Extensible markup language , Metadata , Infrastructure

数字图书馆以传统图书馆为基础，又不同于传统图书馆。随着网络技术的发展、Web 的繁荣，数字图书馆的实践和研究已经成为一个全球性的热点。在过去的几年里，由于互联网技术不断发展，网络用户同时也不断增加，网络信息量同时也以每年10倍的惊人速度增长。但是由于缺乏有效的统一信息资源标准，对于信息的管理及其有效的利用则显得相对落后于信息的增长。

在数字图书馆建设成功之前，书目数据是图书馆提供给用户的最大信息资源，但以往的书目格式(MARC)为图书馆数据资源的整合进入网络流通成为最大的障碍，现在的网络环境以 HTML 为基础构建，其只能描述形式而不揭示内容，所以在 HTML 框架内无法充分表达 MARC 格式所描述的书目数据。可扩展标记语言（XML，eXtensible markup language)不仅可以表示数据，而且可以揭示内容，是一种能够有效表达网络上各种资源信息，为信息的整理、存储、交换、检索提供有效途径的技术。XML 的出现，为数字图书馆的建设和信息资源的利用提供了绝好的技术机会。在这种形势下，笔者提出基于 XML 的数字图书馆技术体系结构。

一、基于 XML 的数字图书馆技术体系结构

数字图书馆的技术体系结构是建设数字图书馆系统的基础，是数字图书馆在网络和计算机技术上的具体实现，是未来信息社会处理、存储和应用数字化信息的基本构架。它的发展目标符合下一代互联网(XML 技术为基础)的发展趋势，具有高度开放、方便易用的特点。

该技术体系结构是以元数据(描述数据的数据)库为核心，以实体对象数据库为基础，通过调度系统可以进行数据的检索、查询、搜索、传输、发布，并具有数据整理系统的输入、整理、格式的转化等多种功能的统一整体。其中元数据库是描述数据形式的基本库，对象数据库是存放具体实体的数据库。用户查询系统界面采用 ASP 技术，其结构如图1所示。

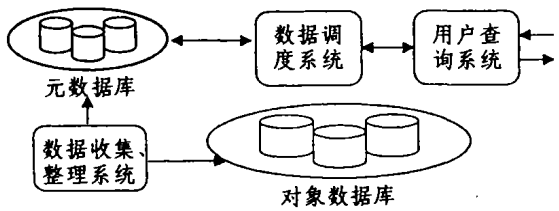

图1　基于 XML 的数字图书馆体系结构

二、数字化资源的加工整理系统

对于建设一个数字图书馆来说，资源的数字化是至关重要的一步。资源加工整理是否好，直接关系到后期图书馆的有效、高效的信息查询和处理，影响到一个图书馆的整体水平。在本系统中着重从两个方面论述资源的加工整理：一是将已有数字资源库转变为 XML 格式的数据；二是对现有大量的以 HTML、PDF 等电子形式存在的资源转变为 XML 格式的数据。下面从两个方面论述。

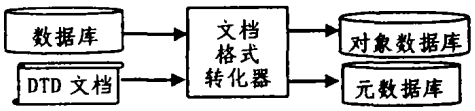

图2　数据转化

1. 现有资源库的 XML 化

90年代以来，随着 IBM 公司数字化概念的提出，各个图书馆纷纷以建立数字图书馆为目标，发展图书馆事业，但是由于当时技术的局限性，近乎百分之百的数字图书馆的资源是以关系型数据库的形式存放的，因此现有资源库的 XML 化是非常关键的。图2是数据转化的模型：首先由图书馆专家给出描述一类资源的元数据，将元数据用 DTD 的形式表示，通过文档格式转化器，分别生成元数据库和对象数据库。

实例1　BOOK..DBF 是一个关于书籍信息的关系型数据库，其基本结构为：

BOOK(title(c),author(c),abstract(c),publisher(c),ISBN(c),price(d),content(entity))；

专家提出的可以描述该数据库的元数据元素有：title , author,abstract,publisher,ISBN,price；以 DTD 的形式描述。通过文档格式转变后生成的对象数据库中存放着文章的内容主体，元数据库则以 XML 的形式存储，为：

```
〈?xml version ='1. 0'〉
〈BOOK〉
 〈title〉JAVA 技术〈/title〉
 〈author〉FRANK〈/author〉
 〈abstract〉这是一部关于……〈/abstract〉
 〈publisher〉电子工业出版社〈/publisher〉
 〈ISBN〉. 7-5635-0422-2〈/ISBN〉
 〈price〉36. 5〈/price〉
 〈content URI="http://dlib. ysu. edu. cn/electronic/book"〉〈/con-
          tent〉
〈/ BOOK〉
```

在 XML 文档中明确地给出了描述一本书的元数据以及对象数据所存放的位置。

2. 电子文档的 XML 化

目前流行的电子文档格式繁多，但以 DOC、PDF、HTML 为主。通过剥离原有数据的附加格式标记，生成纯文本文件。在纯文本文件中搜索出 DTD 所指示的元素，进而形成元数据和对象数据。在文档格式转变中，由于文本文件是一种结构性相对差的数据形式，故而从其中抽取信息的过程，不仅要借助文本自身的标注，而且在某种情况下需要自然语言理解技术。

三、资源调度系统

资源调度系统是通过统一资源标识符 URI(Uniform Resource Identifier)来确定数字图书馆中所有数字资源的规则，建立一个管理所有图书馆数字资源的统一管理系统。在用户向系统发出命令时，资源调度系统首先针对元数据库进行操作，因为元数据库提供了良好的信息资源本身的描述信息，同时具有体积小的特点，故而检索效率高，命中率高；当系统得到用户需要的信息并且发出对数据对象的访问时，资源调度系统会根据在元数据中提供的 URI 从对象数据库中读取数据。

接上实例1：假设用户通过用户查询系统提出查找关于 JAVA 方面的书籍：利用 Microsoft 提供的 XML 包为工具来实现数据的查询(事实上 XML 将可以支持各种各样的应用服务程序)：

```
Set txml=Server. Createobject("MSXML. DOMDocument");
    //在 ASP 中创建 DOMDocument 对象
txml. Load("http://dlib. ysu. edu. cn/electronic/book");
    //在 ASP 上从元数据库中得到一个 XML 文件
curnode=txml. selectNodes(" * JAVA * ") //查找满足条件的节点
curURI=curnode. childNodes(7). Text//获取对象的 URI 值
```

在具体的实现过程中，因为 XML 的后处理特点，浏览器除了可以显示数据之外，还可以分类和过滤数据。当用户再次使用相关数据的检索时，由于前一次元数据可能驻留于浏览器，故不需要从后台服务器读取元数据，从而提高了检索的速度。

四、用户查询和服务系统

在本系统中，数据发送到桌面后，能够用多种方式显示。XML 定义的数据允许指定不同的显示方式使数据更合理、更富有个性要求。本地数据能够以客户配置、使用者选择或其他方式动态地表现出来。XSL 可扩展样式语言(XML Stylesheet Language)为数据的显示提供了技术支持，例如：能够处理元素、属性和内容；能实现条件处理；能封装用户自定义的函数；能实现复杂的、桌面显示的页面布局和样式。用户可以通过多种终端，如 PC、WAP 对数据图书馆进行访问，查询系统界面更加丰富。

图3　XSL 的作用示意图

XSL 处理器如何结合 XSL 转换 XML 文档的示意图如图3所示。一个 XSL 文件片断如下：

```
〈xsl:template match ="BOOK"〉
〈html〉
    〈head〉
      〈title〉〈xsl:value of selset ="title"〉〈/title〉
    〈/head〉
    〈body〉〈xsl:apply - tempates/〉〈/body〉
〈/html〉
〈/xsl:template 〉
〈xsl:template match="BOOK/title"〉
〈h1〉〈xsl:apply - tempates/〉〈/h1〉
〈/xsl:template〉
〈xsl:template match="BOOK/author"〉
〈b〉〈xsl:apply - tempates/〉〈/b〉
〈/xsl:template〉
```

整个样式表中生成了一些模板，在由源 XML 转换为结果时，将对整个 XML 文档节点进行遍历，当遇到一定的匹配模式时(如遇到 BOOK)，就根据其模板规定样式进行转换。完成转换后，就交给具体的输出和显示程序处理。此外，XSL 是对立于输出设备的，其转换目标不仅仅局限于 HTML。

XSL 还规定了其他的指令，如：样式表包含命令 xsl:include 、生成元素命令 xsl:element、循环处理命令 xsl:for、条件处理命令 xsl:if 等，故借助 XSL 可以输出任何满足用户需要的显示格式。

五、数据安全保护系统

目前数据库安全保护机制一般采取在用户名和密码检测的基础上，实现对数据库的保护。采用以 XML 为基础的技术体系结构，对数据的安全控制可以达到任意级别。实现方法有：其一，在需要安全控制的元素前面加上访问者级别和权限说明，这些权限包括检索、读取、排序、统计和修改等；将安全权限赋予一个值，不同的安全级别值表示不同的访问权限，例如，值为1时表示可检索、排序；值为2时表示可读且包含值为1时的操作；值为3时表示在值为2的基础上，增加了写的权限。其二，对于文件规模和系统安全，建立单独的安全控制文件。安全控制有文件的重命名、新建、删除、复制、移动等。

以上面的 XML 文件为例说明一个典型的安全控制元素

```
〈BOOK security-level ="2"〉
 〈title〉JAVA 技术〈/title〉
 ……
 〈price〉36. 5〈/price〉
〈content security = "3", URI = http://dlib. ysu. edu. cn/electronic/
                book"〉〈/content〉
〈/BOOK〉
```

上面的安全控制元素表示，对于 BOOK 节点，拥有权限为2的用户，可以对所有节点进行检索、读操作；权限为3的用户，可以对 content 进行写操作。

在文件级别上，可以对所建 XML 格式的文件进行安全控制，例如：

```
〈security〉
    〈file access〉
        〈reader name="r1", access mode="read"〉〈/reader〉
        〈reader name="r2", access mode="write"〉〈/reader〉
    〈/file access〉
〈/security〉
```

XML 文件在安全控制理论上可以达到任意级别，这对于数据库的安全保护来说，有着重要的意义。

结束语　数字图书馆拥有海量的资源，是数字图书馆的基石，本文提出的基于 XML 技术的数字图书馆技术体系结构，通过以 XML 方式组织、存储数据，并在此基础上探讨关于对象数据库、元数据库的管理、数据的调度、数据的安全保护机制等，对未来数字图书馆的建设有重要意义。