●黄 晨

高可用性数字图书馆框架研究

摘 要 数字图书馆的核心是用户管理和馆藏管理,系统应该建立以存储管理为基础的数字图书馆架构。该架构有 3 个主要特征:基于存储的资产管理、以事件触发的工作流模型、分级权限认证。图 2。表 2。参考文献 6。

关键词 数字图书馆 数字空间 数字对象格式 工作流
分类号 G250.76

ABSTRACT The author thinks that the core of digital library is user management and holding management, and we should build a framework of digital library based on storage management. The framework has three main characteristics: storage-based assets management, event-activated workflow model and stratified privilege authorization. 2 figs. 2 tab. 6 refs.

KEY WORDS Digital library. DSpace. DjVu format. Workflow.
CLASS NUMBER G250.76

随着网络的发展,人们已经习惯于在线检索,无论是已经出版还是未经出版的,是印刷的或是电子的文献。而现状是,无论公司厂商的原型系统或者模型构建[1~2],在大多数的数字图书馆模式中,图书馆的 Web 服务独立于自动化系统,外部用户能够访问的自动化系统数据只有 OPEC,要实现统一定位显然是困难的。能不能有一个既可以把图书馆的内外用户统一,又能够将馆藏对象(Collection)统一的系统呢?换句话说,它能够用同一个机制管理用户,而且能保证印刷和数字文献的统一索引和定位。为此,我们试图突破现有的模式,探讨一种新的数字图书馆框架。

1 系统特征描述

数字图书馆的核心是用户管理和馆藏管理,系统应该建立以存储管理为基础的数字图书馆架构(如图 1)。

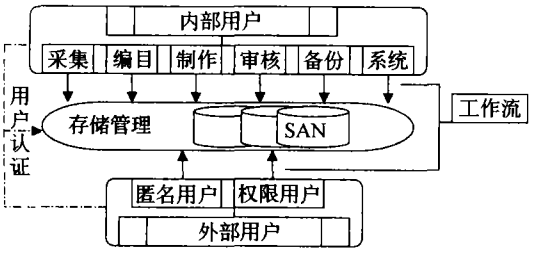

图 1 数字图书馆模型

其特征可以归结为 3 个要点:

(1)基于存储的资产管理;

(2)以事件触发的工作流模型;

(3)分级权限认证。

1.1 基于存储的资产管理

系统首先是基于存储管理的,它应该提供对所有数字格式的支持,比如文档、书籍、数据库、计算机程序、虚拟和仿真模型、多媒体出版物和学习对象等等。另一方面,对于印刷型文献,也可以通过元数据进行存储管理,利用 URL 和馆藏地点来区分文献服务的方式。

数字对象的存储模式有两种:"位"存储(bit preservation)和功能存储(functional preservation)。位存储保证提交的数字材料没有任何改变,即每一位(bit)均保持原样。"位"存储的文件在 5~10 年以后可能不能被当时的用户阅读或执行,但我们假设会有"数字考古学家"(digital archaeologists),在多年以后仍然可以将文件复原出来。功能存储的实现是通过随着时间的变化改变存储内容的格式,以保证当旧有格式淘汰后,数字材料还能够被随时调用(可浏览、可运行、可检索等)。显而易见,功能存储是理想的存储状态,但需要更多的技术和经费支持。

数字对象的存储格式,可以参与 MIME 数据格式规范,定义系统支持的数据格式。考虑到信息技术的发展,也应该允许不支持的格式存储。数据格式可以按 3 个层次定义:支持的格式、知道的格式和不支持的格式(如表 1)。

表1 数据格式的3个层次定义

支持的格式	支持的格式将利用格式迁移技术进行功能存储。
知道的格式	知道的格式意味着系统无法承诺进行功能存储,但是作为一种流行格式,可以尝试通过第三方提供的转换工具完成格式迁移,从而实现功能存储。
不支持的格式	不支持的格式意味着系统没有足够的信息保证功能存储。

有些文件格式很容易向通用的格式转化(像TIFF图像或XML文档),而有些格式会由于专利等原因(像CAD文件),很难实现功能存储。所以系统应该同时支持"位"和"功能"存储。而且在系统的采集模块应能够提供采集转换系统,以将用户提交的材料从"知道的格式"和"不支持的格式"转换为"支持的格式"。

1.2 以事件触发构建的工作流机制

工作流(work flow)在计算机应用中十分常见,但是并没有引入数字图书馆的运作,现有系统仍然以职能部门为基础划分应用模块:流通、期刊、采访等。由于传统图书馆与现代图书馆之间的差别,使得现有的自动化系统根本无法适应数字图书馆的要求。另一方面,国内外的图书馆管理模式不同,就是国内图书馆间也存在较大差异,这就造成国外引进的系统很难在国内完善应用。

工作流的特点是可定制,采用可定制工作流方式,可以随时根据职能变化设置工作流。对于一个特定的工作流只有两种用户:认证用户和非认证用户。认证用户根据其权限策略在工作流中完成相应的职能。另一方面,工作流采用事件触发的启动机制,不同的事件触发不同的工作流,一个工作人员可以根据不同的任务选择进入不同的工作流。

典型的事件如新书采购:书商(外部用户)向图书馆系统提交书目,新书通告被放入采访部(acquisition)的"任务池",采访部的专业人员(内部用户)可以从"任务池"中获取书目进行审核(相应的书目将同时被移除,以免重复相同的工作流)。如果审核通过,系统产生订购单。如果是即时生效的数字资源,这一事件将触发系统的元数据编辑工作流。如果审核不通过,用户将得到附有审核人员意见的通知书。示意如图2。

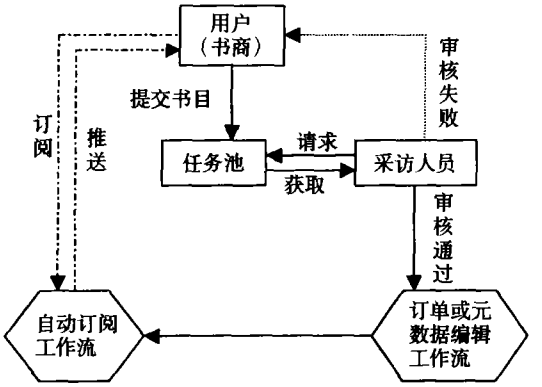

图2 基本工作流

对于数据采集,授权用户可以将自己的材料提交系统,触发馆藏接收工作流。比如教员的课件、学生的学位论文,这些数字文献本来就是数字图书馆的馆藏来源,通过馆藏接收工作流经过专业馆员的审核和元数据编辑,可以很方便地进入馆藏。

1.3 以分级权限控制的管理体系

系统要面对不同的用户请求,某些系统的功能,像检索、浏览等可以是匿名的,但是像提交、订阅以及管理等功能则需要经过用户认证。所以系统应该有可定制的认证系统,采用"缺省否定(default deny)"的策略。比如,一个用户对某个元数据具有"读"权限,但他未必就拥有对此元数据指向的材料的"读"权限。在系统中可以设定的权限如表2。

表2 系统中可以设定的权限

READ读	知道某个对象存在,可以浏览相关的元数据。
WRITE写	修改和某个对象相关的元数据,但不可删除。
ADD添加	允许增加对象,如果需要向一个专题空间提交对象,用户必须拥有这一权限。
REMOVE移除	删除对象。
WORKFLOW工作流	可以加入某个或几个特定的工作流。

权限策略可以针对个别的用户和匿名有户,也可以适用于一个电子工作组或职能部门。

2 核心技术概要

2.1 数字对象格式——DjVu

数字对象格式主要是针对图书馆的原生数字材料(Bron digital marterial)而言的,即图书馆自己进行数字化的书、刊、学位论文等。目前大部分的数字对象格式,或者是由 ASCII 字符流和一些标记(tag)语言比如 HTML/SGML,或者是页面描述(page description)语言像 Adobe 公司的 PDF 所组成。但是当文档具有视觉元素(Visual content)相关性时,上述格式就有了局限。这里所谓的视觉元素包括:不规则字体、纸张的颜色和质地(这对于古籍尤其重要),当然还有表格、图形、公式以及手写体等等诸多方面。为此,我们需要一种适合通过网络发布扫描文档、数字文档或照片中高解析度图像的数字对象格式——由 AT&T 实验室 1996 年针对 WEB 发布开发的图像压缩技术 DjVu(发音为"déjàvu")[3]。

典型的 DjVu 文档压缩率比现有的如 JPEG 和 GIF 格式彩色文档高 5～10 倍,比 TIFF 格式的黑白文档好 3～8 倍[4]。以 300DPI 全彩色方式扫描的文档可从 25MB 大小被压缩至 30～100KB 大小。对 300DPI 的黑白页,经压缩后大小通常变为 5～30KB。高解析度的扫描页可被控制在 HTML 页大小(50KB 左右)。对同时存在图片与文本的彩色文档,在相同质量情况下 DjVu 比 JPEG 格式小 5～10 倍。

需要特别指出的是,DjVu 不仅仅是一种图像压缩技术,而是像 PDF 那样的内容格式(Content Formmat)。它拥有完整的导航、搜索、超链接等可编程模块。并且更适合于彩色文档,尤其是图文混排文档的发布和传递。

DjVu 还支持隐含文本(hidden-txt),即前端以图像方式体现文档原貌,后台集成文字信息。一个包含隐含文本的文件在被浏览时,浏览器上的文本选择按钮会被激活。这时按下文本选择按钮,用鼠标在图像中相应文本内容上画一个矩形框,被选中的部分即反色。被选中的文字可以被复制到系统的剪贴板中,相应还可实现对图像中文字内容的搜索、查询。

DjVu 是一个公开的标准,文件格式规范,解码器实现及部分译码器是开放的。它有适用于各种操作系统的浏览器插件程序,DjVu 插件允许对文档方便地浏览及缩放。DjVu 提供了扫描文档与数字文档的统一发布平台,而且效率极高。考虑到它支持将 PDF、PPT、HTML 和 DOC 资源转换成 DjVu 格式,这就保证了我们可以充分利用已有的数字资源。

2.2　数字对象存储－SAN

在存储技术不断发展和转变的过程中,新的存储需求逐渐被用户提出来。但总的来说,用户要求存储的架构必须可以根据需求来进行扩展(Scale-out)和放大(Scale-up)。其关键就是要在不影响或改变原有操作和存储环境特性的前提下,轻松地实现存储能力的增强。

SAN(存储局域网)是以数据存储为中心,采用灵活的网络拓扑结构,通过具有高传输速率的光纤信道的连接方式,以传统 SCSI 协议传输数据的一种体系。SAN 有许多的优点:(1)SAN 可以无限扩展,用户可以灵活地在 SAN 结构上添加各种存储设备,整个系统扩展起来很灵活;(2)由于采用光纤信道的传输方式,SAN 上的传输速率非常高,当前的传输速度为 200Mbit/秒;(3)SAN 的数据传输采用块传输方式,可以完成大规模数据传输;(4)SAN 独特的结构特别适合于各种备份,实现现在比较流行的 Lan-free 和 Serverless 的备份方式[5]。

数字图书馆对于数据存储容量的需求几乎是无穷的,因此采用存储局域网架构和虚拟存储技术是必须的。这一模式有良好的稳定性、扩展性、安全性和可管理性。能够为数据共享、数据集中、综合灾备等内容提供可靠的保障。

2.3　数字对象管理——DSpace

数字对象管理是在存储管理的基础上进行的,SAN 的存储管理软件主要针对于存储设备,而数字对象的管理是针对于内容的。MIT 和 HP 实验室合作两年多研究的 DSpace(数字空间)平台是一个基于 BSD 开放源代码许可的管理系统,从 2002 年 10 月开始在 MIT 正式服务。

DSpace 是一个专门的数字资产(Digital assets)管理系统,它管理和发布由数字文件或"数字流"(bitstreams)组成的数字条目(item),并且允许创建、索引和搜索相关的元数据以便定位和存取该条目。它包括的基本概念主要有:

数字空间群(DSpace communite):数字空间是针对数字材料的长期保存而设计的。数字材料来源于不同的组群,比如大学的院系、实验室,图书馆的采访、编目等部门,这些依据不同的授权完成不同任务的组群称做数字空间群。

电子用户(e-people):DSpace 的用户可能是教师、学生,也可能只是一些计算机系统,因此,DSpace 把用户称做"E 人"。

工作流(workflow):DSpace的运作方式、数字材料和元数据在被接受之前经过事先设定的流转审核的步骤。

信息订阅(subscription):用户可以向DSpace发送订阅请求,以便在新材料到来时收到带有内容提要的电子邮件提示。

如上所述,DSpace是一个数字资产管理系统,本身不是为数字图书馆而开发的,但它完全可以被引入到新一代数字图书馆框架中。

3 构建基于DSpace的数字图书馆

DSpace是一个开放源代码的软件平台,可以运行于所有UNIX系统,像Linux或者HP-UX等。它对应于数字图书馆的5个技术环节:数字资源采集、数字对象存储与管理、搜索技术、信息传递技术和权限认证[6]。我们可以根据自己的需要来修改和扩展它的功能。由于Java虚拟机是由HTML和部分嵌入的Java代码组成的,基本上不需要触动其核心的代码就可以方便地修改它。

数字资源采集是数字图书馆的对象来源,包括创建原始内容和捕获开放系统中现存的资源。DSpace不支持原始内容的创建,但可以以工作流方式捕获任何支持主动文档开放协议(Open Archives Initiative)的数字资源和元数据,同时也通过这一方式完成信息传递。

DSpace可以存储、管理和发布任何已经和未经出版的本地馆藏,保证印刷和数字文献的统一索引和定位。采用如前所述的存储局域网体系加虚拟存储技术,实现用户与存储的双向透明。在此基础上通过DSpace制定完善的用户分级权限认证策略,利用消息驱动的工作流机制进行数字图书馆运作和服务。

检索是DSpace挖掘(discovery)的重要组成部分。用户对于搜索引擎的要求是很高的,所以DSpace的目标就是提供尽量多的检索特色。DSpace的索引和检索模型有一个API接口,允许非常方便地索引新内容,重建索引以及在指定范围内检索。这个API来自免费的Java搜索引擎——Lucene。Lucene支持字段检索、停词(stop words)、词干(stemming)以及不重建索引增加新的索引内容的能力。

所有的DSpace用户界面都是基于Web的,并且包含一系列可定制的支持权限认证的界面:用户提交界面、搜索界面、系统管理界面以及提交审核的各种界面。这使得DSpace的使用和管理都十分方便。

在虚拟存储技术的支持下,利用DSpace联盟,还可以实现虚拟馆藏:联盟的馆藏对于用户完全透明,用户通过Web界面访问元数据,并发送资源请求。DSpace可以根据标准协议进入不同院校的系统,比如康奈尔大学的arXiv,通过本地的存储过程自动获取有关的文档备份。用户只需点击锁定(located)的条目,它就会被下载到Web浏览器中。如果是一种Web支持(Web-native)的格式,它就会立刻显示在浏览器中;否则它就被存储在用户的计算机上,例如Excel数据表或者CAD文件,需要以特定的程序来打开它。

浙江大学图书馆一直关注着新一代数字图书馆的发展,并致力于高可用性的数字图书馆框架研究。我们认为,围绕DSpace核心,制定完善的策略,设计合理的工作流,采用成熟的工业标准和协议,构建一个高效、稳定、易用的新一代数字图书馆是完全可能的。我们希望能够在较短的时间内公开这一原型系统,以供交流和完善。

参考文献

1 杨宗英,郑巧英.再论数字化图书馆的现实模型——兼谈上海交大计划建立数字化图书馆的雏型.现代图书情报技术,1997(1)
2 郑巧英,杨宗英.台湾的图书馆自动化和数字图书馆建设.大学图书馆学报,1998,16(5)
3 http://www.djvuzone.org/djvu/djvu/djvuspec/001.djvu
4 L. Bottou, P. Hafther, P. G. Howard, P. Simard, Y. Bengio. and Y. LeCun. High quality document image ompression with djvu. Journal of Electronic Imaging,1998,7(3)
5 黄晨.存储区域网模式前沿技术研究.图书馆杂志,2001(4)
6 董慧,安璐.数字图书馆关键技术的分析与启示.情报学报,2002(6)

黄 晨 浙江大学图书馆数字资源建设中心主任,副研究馆员。通讯地址:杭州。邮编310027。

(来稿时间:2003-11-10)