文章编号:1001-5132（2007）04-0498-05

通用数字图书馆的仓储管理方案研究

杨学明

（宁波大学 网络中心，浙江 宁波 315211）

摘要：在数字图书馆时代，如何保证数字资源的长期可用性和服务的稳定性，对数字资源进行有效的管理和服务，是当前数字图书馆研究中的一个热点前沿话题. 在研究分析了国外数字图书馆仓储管理的几种典型解决方案后，提出了一种基于数字对象的通用数字图书馆仓储管理方案. 结果表明：该方案实现了统一存储各种属性不同的数字资源，并提供了很好的资源管理和访问接口.

关键词：数字图书馆；数字对象；数字仓储；数字资源

中图分类号：TQ171 文献标识码：A

随着信息技术的迅猛发展，越来越多的数字资源需要存储管理在数字图书馆中. 这些数字资源具有 4 个方面的特征：(1)资源类型的多样性；(2)元数据规范的差异性；(3)资源在网络上的分布性；(4)提供资源服务的特殊性. 由此产生了相关研究问题为：(1)如何将这些属性各异的数字资源高效统一地存储管理起来；(2)如何基于不同的资源提供不同的服务. 解决这些问题的关键是如何构建 1 个通用的数字仓储管理系统. 数字仓储是数字内容的存储管理技术，它能够在同 1 个存储体系下容纳各种类型和众多复杂格式的数字内容，并为目标用户提供对这些数字内容长期稳定服务，该技术被广泛应用于创建信息环境、出版系统和长期保存系统. 目前国内外已经提出了许多具有实用性的数字仓储解决方案，如 MIT 的 DSpace 系统、Cornell 大学与 Virginia 大学图书馆合作的 Fedora 系统和 California 大学的数字仓储系统等.

1 国外数字资源平台简介

1.1 Greenstone

Greenstone 是一组有关数字图书馆建设的开源软件，它提供了在网络或者 CD 中组织和发布信息的 1 种新方式. 它是新西兰数字图书馆计划的一个部分，并且得到了联合国教科文组织和 Human Info NGO 的协助. Greenstone 是开源软件，可以在 Linux、Windows、Macos 3 种平台上运行，现在此软件推出了 3.0 版本，用 Java 重新撰写了所有模块，功能更为强大.

1.2 Fedora

Fedora 为通用的数字存储项目，由弗吉尼亚大学图书馆和科内尔大学研制. 它利用网页技术进行分布式数字信息系统管理以及提供相关服务，比如 XML 技术以及其他技术. 系统默认采用的标准是都柏林元数据集. 通过元数据，可以进行 OAI

收稿日期：2006-11-06. 宁波大学学报（理工版）网址：http://3xb.nbu.edu.cn
基金项目：浙江省社科规划课题（NM05GL02）；宁波大学科研基金（XK0600083）.
作者简介：杨学明（1978－），男，浙江湖州人，硕士/工程师，主要研究方向：软件工程. E-mail: yangxueming@nbu.edu.cn

元数据采集.

1.3 DSpace

DSpace 是数字保存系统，也是专门的数字资产(Digital Assets)管理系统，其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. 此项目于 1999 年由 MIT 和 HP 实验室合作研究开发的，到 2002 年 10 月，该平台开始在 MIT 正式服务. 此后由 MIT 图书馆和 HP 实验室一起向全世界公开了基于 BSD 开放源代码许可的 DSpace 源码. DSpace 的主要代码均用 Java 编写，可运行于所有 Unix 系统，如 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. 此系统主要用于某一组织机构采集、加工、保存本单位的研究成果.

1.4 DLXS

DLXS 是密西根大学数字图书馆推广服务 The University of Michigan Digital Library eXtension Service 的缩写. 它为教育机构和非赢利机构提供了完整构建数字图书馆的框架和基础，特色在于搜索引擎以及一些基于类的中间件，由其提供的中间件非常有用.

1.5 LOCKSS

LOCKSS 项目源于 Sun 公司与斯坦福大学之间的一项合作，他们共同创建了 LOCKSS(Lots of Copies Keeps Stuff Safe)系统，主要是为了解决电子出版物的收集和永久性保存问题. 基于 Java 技术的 LOCKSS 系统是开放性源码的分布式系统，它无需中心级管理就能运行在一些廉价的 PC 机上. 斯坦福大学图书馆的 LOCKSS 系统就运行在一个由廉价计算机组成的网络上，它可以监控计算机硬盘中存储的所有文档. 在测试运行中模拟了一些出版者人为的错误，测试结果表明：如果一些文件被删改或毁坏，自动缓存系统就会用完整的文档来取代它们.

2 数字对象框架

数字对象框架(Digital Object Architecture)最早是由 William Y 于 1997 年提出的，数字对象是数字对象框架的核心，其结构如图 1 所示. 从面向对象的角度来讲，数字对象(Digital Object)是一个唯一标识的网络实体，它用来封装和描述属性不同的数字资源，并且提供访问数字资源的机制.

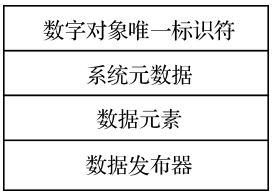

图 1　数据对象结构

数字对象由数据和数据发布器组成. 数据包含 2 部分的内容：核心元数据(Key Metadata)和数据元素(Data Element). 核心元数据又包含 2 部分元素：数字对象唯一标识符(Digital Object Identifier)和系统元数据(System Metadata). 数字对象的唯一标识符用来唯一地标识一个数字对象，以便于在命名空间内唯一地引用该数字对象. 而系统元数据用来描述整个数字对象，用来管理该数字对象和建立数字对象的索引.

数据元素是数字对象所包含的数据，它可能是元数据，也可能是数据本身. 如果是元数据，可能是各种格式的，比如 Dublin Core，MARC. 如果是数据本身，可能是文本、图像、音频和视频数据格式. 两者在数据元素的层面上是没有区别的. 也就是说，数据元素将数据和元数据统一对待，这样就可以统一处理资源类型格式不同和元数据格式不同的数字资源. 1 个数字对象包含 1 个或者多个数据元素. 数据元素是数据发布器中实际操作作用的对象，它可以位于本地机器上，也可以位于远程机器上. 当位于远程机器上时，需要能够通过某种方式比如 HTTP 协议或者 FTP 协议访问到该数据元素. 这样，数字对象包含的数据可以分布在网络中的不同位置.

数据发布器(Disseminator)是数字对象内部的一种结构，对应着一种发布数字对象内容的方式. 1 个数字对象有 1 个或者多个数据发布器. 1 个数据发布器含有数据内容类型(Content Type)和操作的数据元素. 数据内容类型对应着一系列的数据内容类型操作接口(Content Type Signature)和数据内容类型操作(Content Type Implementation)的实现. 数据内容类型，也就是操作接口和操作的实现本身，是独立于数字对象而存在的. 操作的数据元素就是操作的实现将要作用的数据元素，它可以是数据，也可以是元数据.

3 通用仓储框架

仓储(Repository)是网络上的存储系统，它为数字对象的存在提供了容器. 仓储实现了数字对象的存储管理，并且通过一定的访问控制策略提供了数字对象内容发布的机制. 外界通过仓储访问协议(Repository Access Protocol)来管理和访问数字对象. 数字对象存放在仓储里面，1 个数字对象对外部可见的只是它的唯一标识符，与数字对象的交互只能通过仓储访问协议进行. 为了使仓储系统能够按照需求进行更新或升级，增强系统的可维护性和可扩展性，我们设计的仓储框架采用了多层次结构，将系统框架分为 Web 服务层、会话层、中间层和数据存储层，如图 2 所示.

Web 服务层分别提供了基于 HTTP 和 SOAP 协议的管理接口、访问接口、查询接口，其体系结构都构建于 Web 服务技术之上. 管理接口定义了管理仓储的公开接口，包括创建、修改、删除数字对象或者数字对象的数据元素(包含数据和元数据). 管理服务模块和底层的模块交互，进行数字对象及其数据元素的读取和写入. 管理服务模块提供了 1 组抽象的操作允许客户端操纵数字对象，而不必关心底层的存储格式及其存储媒体等问题. 访问接口定义访问数字对象的公共接口，访问数字对象的方法包括请求执行 1 个数字对象数据发布器中的方法. 为了发布数据，底层的模块需要动态地将数字对象的操作方法和数据元素进行绑定，动态地实现数据的发布. 查询接口则提供了资源的内容和索引查询.

会话层主要功能是对访问用户进行统一认证，可以通过 IP 地址受限和 HTTP 验证等方法. 通过统一认证后，用户根据自己在系统中的角色，获得不

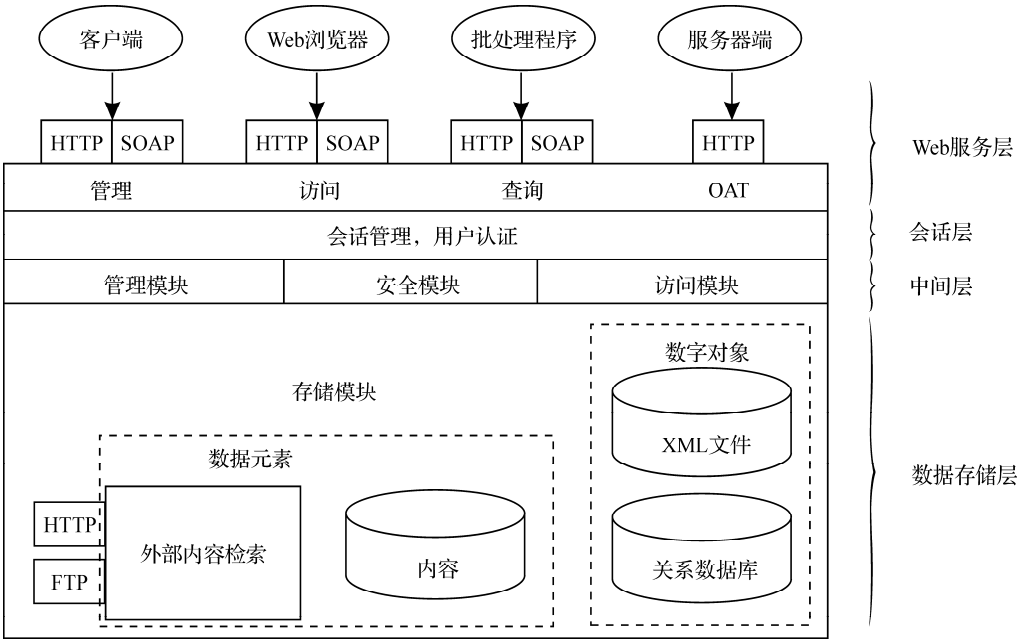

图 2 通用数字仓储框架结构

同的权限，可以访问允许的内容和操作.

中间层是内部服务层，包含管理模块、安全模块和访问模块. 管理模块和访问模块分别实现了管理和访问接口. 管理模块主要管理数字对象的操作、对象完整性校验以及数字对象统一标识符的生成；安全模块实现系统用户和仓储的安全策略；访问模块实现数字对象的传播和数字对象行为的映射.

数据存储层，作为仓储框架中的底层核心，主要进行保存数据和对数据进行读、写和删除操作，这里所指的数据不仅指数据流还包括数字对象的元数据包文件. 在存储模块中，每个数字对象在保存数据流的同时，以 XML 格式封装元数据生成元数据包文件. 对于每个对象的元数据而言，系统采用直接封装的方式，这种方式一方面是便于整体管理对象的元数据，另一方面也在一定程度上支持了系统的开放获取.

4　具体应用

4.1　创建数字对象

在通用框架中，数据对象都是以 XML 文件的格式存储和交换，并采用 METS(Metadata Encoding and Transmission Standard)编码. METS 用来将数字图书馆中数字对象相关的描述性元数据、管理性元数据和结构性元数据进行编码的一个标准，采用 W3C(World Wide Web Consortium)的 XML Schema 语言表达. 该标准由美国数字图书馆联盟 DLF(Digital Library Federation)开发，由美国国会图书馆的网络发展和 MARC 标准办公室负责维护.

目前系统各个模块内部采用 Schema 驱动生成用户界面的技术，多个模块的数据在完成元数据标注的时候，组合成完整的 METS 文档，存储在服务器端. 采用 METS 格式编码的 XML 文件中包括 METS 文件头(MetsHdr)、描述性元数据(DmdSe)、管理性元数据(AmdSec)、文件组(FileSec)、结构图(StructMap)、结构链接(StructLink)和行为机制(DehaviorSec) 7 部分内容. 系统的著录工具提交标准 METS 编码的 XML 文件，标准 METS 格式利用文件组(FileSec)和结构图(StructMap)实现对图书等资源的复杂结构的保存和揭示. 其中：文件组部分包含 1 个或者多个<fileGrp>元素，每个<fileGrp>中是构成该数字对象的某种资源类型的文件列表. <fileGrp>中包含 1 个或多个<file>元素，分别关联实际对应的文件. 结构图部分通过一系列嵌套的 元素来表示分层的结构，每个 元素记载了该层次的类型信息，同时包含若干个指向内容文件的指针<fptr>元素，<fptr>指针用来指定该 中包含的内容文件在<fileGrp>中对应的部分，实现结构和物理文件之间的关联.

具体实例如下所示：

```
<METS:mets TYPE="GeneralObject" LABEL="数字图书馆"  ID="nbulib_sztsg" PROFILE="MATHSTEB" >
……
<METS:fileSec>
 <METS;fileGcp ID=”DATASTREAMS">
  <METS:fileGip ID ="DSl" STATUS="A">
   <METS: file CREATED = "2006-09-05T15:32:15.112+08:00" MIMETYPE="application/teb" STATUS="A" ONNERID="M" ID="DS1.0">
    <METS: FLocat LOCTYPE="OTHER" xlink:href="nbulib_sztsg.teb"  xlink:title="数字图书馆"/>
……
 </METS:fileSec>
</METS:mets>.
```

4.2　虚拟馆藏

虚拟馆藏被定义为一组数字对象的聚合体，有着自身的资源展示方式. 这些数字对象物理地存在于数字图书馆系统的仓储中，而逻辑上属于虚拟馆藏. 虚拟馆藏非常灵活，它可以支持各种数字资

源提供的各种服务，可以支持二次开发，具有很好的可扩展性. 在通用框架中，我们采用 XML 格式的配置文件，并定义相应的 Schema，保存虚拟馆藏工作所必须的各种信息：(1)虚拟馆藏的基本属性，即虚拟馆藏的名字，它包含哪些数字对象. (2)虚拟馆藏中的数字对象所建立的一个或多个索引，用来支持一个或者多个检索服务. (3) 1 种或者几种针对虚拟馆藏中数字对象的检索方式，检索针对外建索引进行.

用户可以定制配置文件中的索引信息，来控制索引器的工作；也可开发自己的索引信息抽取器类或者索引器类，为新的资源实现新的索引. 通过对虚拟馆藏中的数字对象建立索引，可实现全文检索服务，或者基于特征的图片检索服务等各种服务.

5 结语

本文设计并实现了通用数字图书馆仓储管理框架，解决了数字图书馆领域中统一存储管理各种属性不同的数字资源和基于属性不同的数字资源提供各种服务这两大难点问题. 该框架具有很好的通用性、可扩展性和互操作性. 采用了以数字对象框架实现底层的资源管理，实现了统一存储各种属性不同的数字资源，并提供了很好的资源管理和访问接口.

参考文献：

[1] 董慧, 安璐. 数字图书馆关键技术的分析与启示[J]. 情报学报, 2002(6):700-707.

[2] 黄晨. 存储区域网模式前沿技术研究[J]. 图书馆杂志, 2001, 20(4):30-31,40.

[3] Brian F L. The open archival information system reference model: introductory guide, digital preservation coalition[EB/OL]. [2004-08-12]. http://www.oclc.org/research/staff/lavoie/lavoie_pubs.pdf.

[4] 高文, 刘峰, 黄铁军, 等. 数字图书馆－原理与技术实现[M]. 清华大学出版社, 2000.

[5] Arms W Y. An architecture for information in digital libraries[EB/OL]. [2007-10-22]. http://www.dlib.org/dlib/february97/cnri/02arms1.html.

[6] Altman M. A digital library for the dissemination and replication of quantitative social science research[J]. Social Science Computer Review, 2002, 19(4):450-487.

[7] 董丽. METS 元数据编码规范及其应用研究[J]. 现代图书情报技术, 2004(5):8-12.

A Study of Universal Digital Library Repository Management Scheme

YANG Xue-ming

(The Network Center, Ningbo University, Ningbo 315211, China)

Abstract: How to ensure the long-term availability of digital resources and the service stability has been drawing more and more research attention in the forefront of digital library. Based on the analysis of a typical foreign digital repository management schemes, an object-oriented universal digital library repository management scheme is proposed in this paper.

Key words: digital library; digital object; digital repository; digital resource

CLC number: TQ171 **Document code:** A

（责任编辑 章践立）