●林 颖 张智雄

一种基于开放源码的数字保存系统设计

摘 要 在数字保存的技术体系上提出一种设计思路:基于开放源码软件实现一个数字保存系统。以 DSpace 为基础进行数字保存系统的扩展,扩展了保存管理功能、摄入功能、仓储功能、存储和访问功能。表 1。图 4。参考文献 7。

关键词 数字资源保存 保存系统 开放源码 技术方案

分类号 G250.76

ABSTRACT The authors propose a design of a digital preservation system based on open source code software. They also use DSpace as the basis for the extension of digital preservation system, which extends the preservation management function, capture function, warehousing function, storage function and access function. 1 tab. 4 figs. 7 refs.

KEY WORDS Digital resource preservation. Preservation system. Open source code. Technical design.

CLASS NUMBER G250.76

所谓数字保存系统,是一个"能够存储各种类型数字资源的场所,并且有一套基本的原则了解这个场所需要存储什么和提供什么样的服务"[1]。而开发数字保存系统是一项大型的软件工程,开发成本高,周期长,维护困难。如何实现数字保存系统是当前一个重要的研究课题。

开放源码软件是指允许任何人免费(或少许收费)使用、拷贝、修改、发布的软件。对这类软件,用户有使用、修改、复制的自由。在图书情报领域,一些有代表性的数字图书馆开放源码软件,如 Greenstone,SPT(Subject Portal Toolkit,学科信息门户),DSpace 等,不断被行业用户熟悉并应用。利用开放源码软件进行系统开发是一种有效的方法。本文通过对各种相关开放源码软件分析和调研,提出一种基于 DSpace 的数字保存系统的实现框架。

1 技术体系

数字保存系统的技术体系应遵循 OAIS 参考模型(开放档案信息系统,Open Archival Information System),但是 OAIS 是一个概念模型而非系统设计模型[2]。数字保存系统不必照搬它,而是应该根据不同的保存需求进行系统功能模块的分解和设计。

在 OAIS 参考模型的指导下,数字保存的技术应该分属于保存管理、摄入、仓储、存储管理和访问 5 个功能块:保存管理主要涉及保存技术策略的选择、保存规划管理、保存工作流管理、保存媒体迁移等方面的相关技术;摄入主要涉及在摄入之前和摄入过程中对数字对象进行规范处理的各种技术,主要的技术有格式标准、格式迁移、格式规范和格式注册技术,信息封装技术,安全检测技术,完整性校验技术和数据功能校验技术;仓储主要涉及在数字对象摄入之后如何对数字对象及其元数据进行管理的技术,主要包括信息模型的构建、保存元数据体系、保存标识体系、内容管理、元数据管理、索引等方面的技术方法;存储主要涉及如何构建大规模安全存储体系,对存储对象进行备份和恢复的技术,主要包括常见的磁带存储、光盘存储、磁盘阵列存储,也包括各种类型的分布式文件系统、基于 NAS 或 SAN 模式的网络存储和基于网格的存储体系,也包括相应的备份和恢复系统;访问主要涉及如何使仓储的数字对象能够被安全方便地访问的技术,主要包括检索浏览技术、基于保存标识的定位技术、认证和授权技术、与第三方的互操作技术。数字保存的技术体系如图 1 所示,本文尝试在这样的技术体系指导下设计一个具有现实可操作性的数字保存系统。

2 系统设计思路

基于已有的或开放源码的系统进行系统构建,是高效设计和实现数字保存系统的重要思路。例如荷兰国家图书馆的数字信息存档系统 DIAS 基于 IBM 的 DB2,Content Manager,TSM(Tivoli Storage Manager)和 Business Objects;Cornell 大学的通用仓储系统 CDS 基于 ENCompass,ArXiv 和 Project Euclid;佛罗里达图

书馆自动化中心的黑色存档系统 DAITSS 基于 MySQL;加利福尼亚大学数字保存仓储 DPR 基于 SRB(Storage Resource Broker),Shibboleth 和 MySQL 等。这些数字保存系统都建立在已有的一些系统之上,极大地提高了开发效率,节省了时间和成本。本文尝试在开放源码软件的基础上设计和实现一个数字保存系统。通过对长期保存资源特性的分析,本文认为数字保存系统可以在内容或数字资产管理系统的基础上进行构建,尤其可以通过开放源码的机构知识库软件来实现数字保存系统。

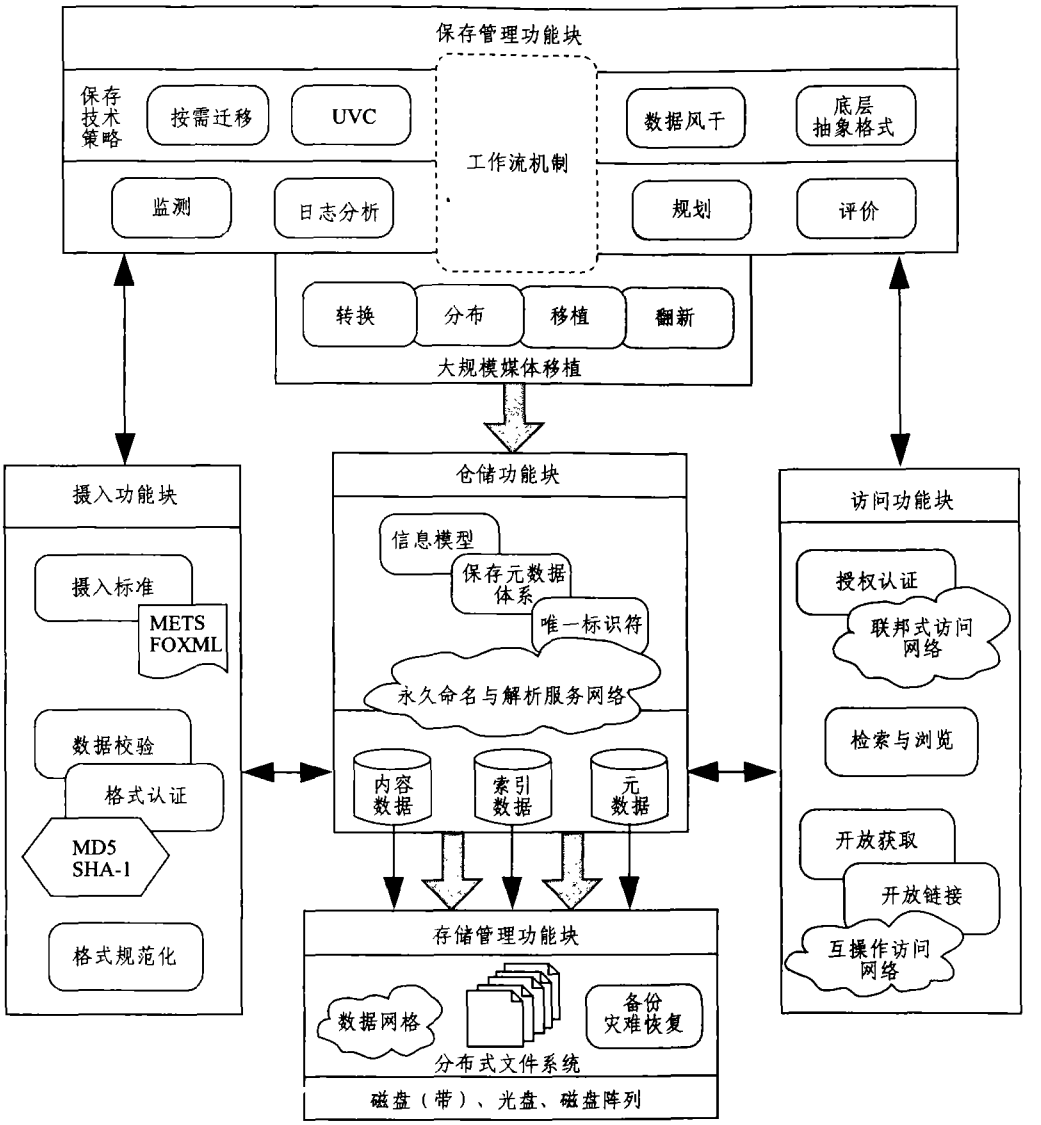

图 1　数字保存系统的技术体系

国外有许多开放源码的机构知识库软件,主要有 DSpace,Fedora,EPrints 等,而且一些研究组织也会不定期有针对性地分析影响较为广泛的仓储并作出综合评价。笔者从管理结构、实用效果、扩展前景等方面进行了分析比较,最终选择以 DSpace 为基础进行数字保存系统的构建。

2.1　DSpace 与 OAIS

DSpace 是美国麻省理工大学与惠普公司合作开发的数字资产管理系统。该系统的目标是创建一个数字仓储用于获取、存储、索引、保存和访问研究组织的数字知识资源。DSpace 遵循了 OAIS 参考模型,基本涵盖了 OAIS 的六大功能模块:摄取、存档、访问、

数据管理、保存计划、管理。OAIS 分解如图 2 所示。

事实上，DSpace 的主要设计目标之一就是实现数字信息的长期保存，它已经具备了许多数字保存的特性，这也就使得基于 DSpace 设计一个数字保存系统有了可行性和便捷性。

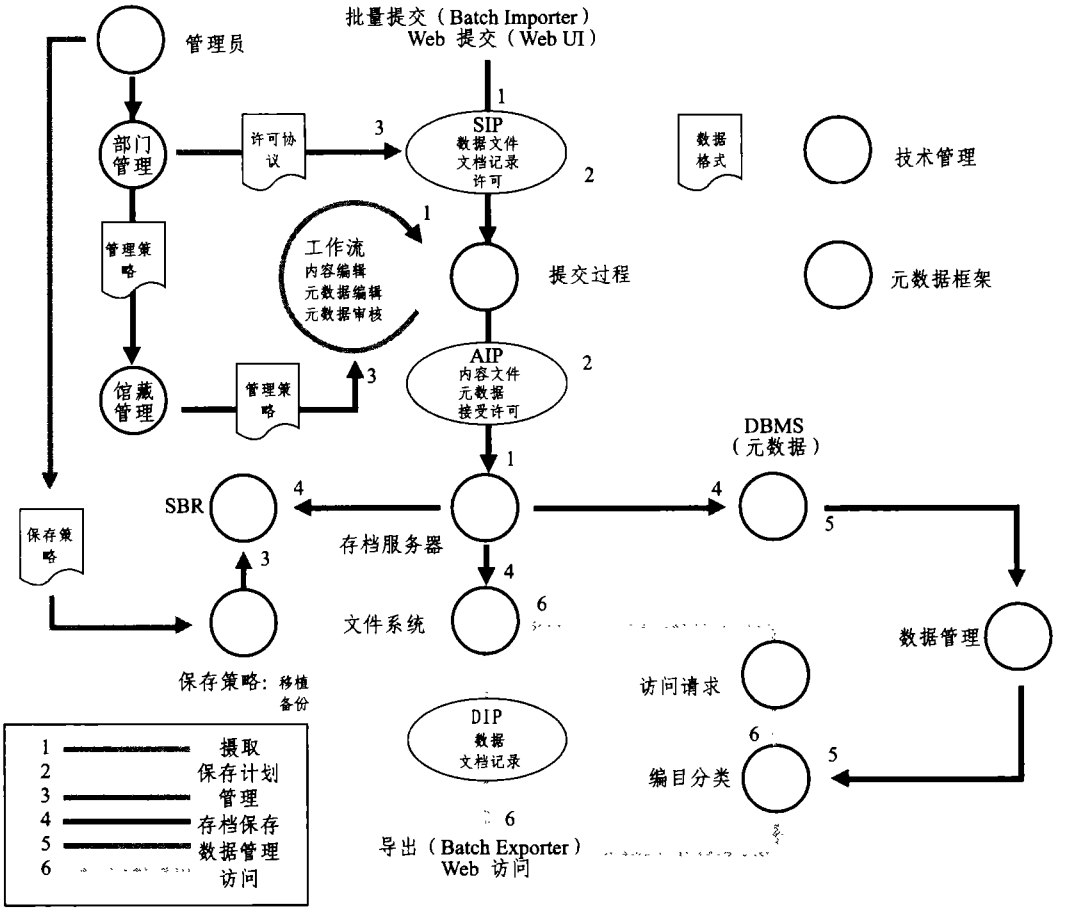

图 2　DSpace 的 OAIS 分解

DSpace 还有灵活的适应性和高度的可扩展性。从系统结构上看，它模块化程度高，具备很强的扩展性。从应用服务集成上看，它集成了多种外部服务，例如 log4j[3] 日志服务、lucene[4] 的索引服务等，而且还致力于实现数字资源之间的互操作服务，例如开放链接等。从标准协议上看，它的技术标准大多都遵循了相关的国际通用标准，除了 OAIS 参考模型，它还支持元数据收割协议 OAI-PMH2.0，DC 元数据格式等。

2.2　DSpace 扩展

DSpace 这些优秀的特性都为本文所主张的基于已有的或开放源码的进行数字保存系统的建设奠定了基础，但它毕竟只是一个机构知识库软件，不能等同于数字保存系统。虽然它可以实现一个信息交流环境，但并没有满足数字资源在长期保存上的需求，

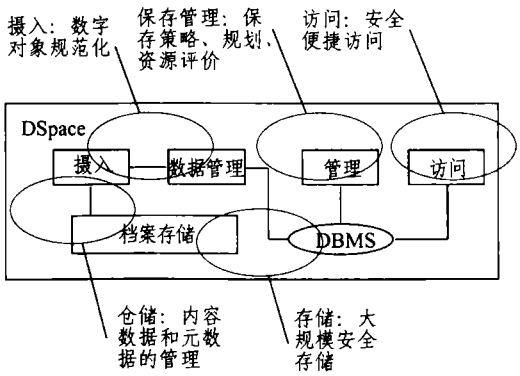

图 3　DSpace 的扩展需求

比如在存储性能表现上的不足,缺乏文件格式的保障机制等。如果要以它为基础构建一个数字保存系统,那么在技术体系的5个功能上都要进行一定程度的扩展,如图3所示。

2.2.1　保存管理功能

数字资源常常会因为技术和环境的变迁而变得无法访问,这对长期保存而言是致命的,制定灵活的保存技术策略并有效管理就成为数字保存系统的重要任务。最常见的是为DSpace扩展多重备份和媒体迁移的保存技术策略,但仅仅这样还不足以保障数字资源长期有效使用,还需要根据DSpace在数据结构和存储结构上的特点,扩大系统在按需迁移(Migration on Request)[5]、仿真、环境封装、通用虚拟计算机(UVC)[6]等方面的技术应用。

数字保存是一个漫长过程,数字保存系统无法回避在何时、采用何种方案对特定(类型)的数字对象、仓储系统和存储系统进行更新和升级,因此DSpace还需要扩展保存规划和系统资源评价等方面的技术应用,比如日志分析技术,这样就能在保存系统实施过程中不断地修订更新完善相关的保存管理功能。

2.2.2　摄入功能

DSpace在摄入功能上只是简单地实现了数字资源的获取(包括元数据),并没有对数字对象进行规范化处理。还需要从3个方面扩展。

首先是格式认证。保存系统要选择适当的文件保存格式以保证文件格式的持久生命力。并非所有的摄入资源都能符合保存文件格式的几大原则,还需要通过其他途径保障这些文件格式的长久有效性。一种就是文件格式的注册,即登记文件格式的详细信息,确保这些文件格式的持久有效和永久可解析;另一种则是文件格式软硬件环境的保存,以便在将来文件格式不可用情况下及时复原当时的信息环境。

文件格式认证后还要对数字资源进行校验,目的是确保传递前后数字资源数据的完整、有效和一致。例如对于一个Doc文件,还需要完成3个方面的验证:确认它是否确实是Doc格式,它是否包含Doc格式的特征内容,其内容是否完整有效。这样才能保证摄入的数字对象在当前是可用的,这也是数字资源在将来具有可用性的基础,一个不完整的数字资源不可能在长久的将来还能正常使用。

数字保存系统还应该尽可能地将不符合保存格式的文件格式进行规范化和标准化,即格式迁移。尤其对于一些私有格式,在数字保存仓储中应该尽可能地迁移为系统支持的文件格式,减少将来因为格式退化而引发的数字资源不可识别。而且,保存文件格式也会随着信息技术的发展而变化,这同样需要格式迁移技术的支持。

2.2.3　仓储功能

DSpace作为仓储软件,在数字对象的内容数据和元数据的管理上有显著优势,比如它的数据结构支持任意格式数据内容的存储,支持含有66个元素的保存元数据集,支持元数据索引等。但它还缺少保存的标识体系和内容索引。

保存标识体系是通过命名标准与解析服务器为数字资源建立一种永久的逻辑标识来实现全球唯一的准确定位,特别是要支持数字资源标识的扩展性、安全性、多重实例、国际化。

为了加快数据的检索过程,提高资源的访问速度和使用效率,内容数据的索引也相当重要,尤其要对不同类型的数据内容建立合适的索引文件,包括多种语言的文本索引、图像索引等。

2.2.4　存储功能

DSpace在存储结构上是关系数据库和文件系统相结合,前者保存元数据,后者保存内容数据。文件系统在存储安全性、存储效率、分布式存储等方面都略显薄弱,还需要构建一个大规模的安全存储体系,比如磁带存储、光盘存储、磁盘阵列存储等,尤其是基于NAS或SAN模式的网络存储和基于网格的存储体系,以其良好的扩展性和强大的功能性满足了用户对存储的需求,可以实现数字资源异构异地的无缝透明存储。

2.2.5　访问功能

在数字保存系统的技术体系中,这一部分需要包括检索浏览技术、认证授权技术、互操作技术等。DSpace通过Web用户服务对外提供了检索浏览功能,但是它的认证授权仅是通过安全套接层协议(SSL,Secure Socket Layer)对用户的名字和口令进行验证。随着保存系统协作交流服务的开展,这种认证授权会逐渐无法保证访问的安全性和便捷性,因此需要以单点登录的方式进行用户身份管理,通过安全上下文或凭证,在多个应用系统之间实现(进行)传递或共享信息。

DSpace基于一个开放系统的要求应用了一些开放协议规范,比如OAI元数据收割。但是随着互操作技术的发展,更多的开放协议需要集成到保存系统中,比如RSS数据收割、开放链接(DSpace只支持与SFX开放链接系统的集成)等。

3 系统实现框架

为了实现上述5个部分所讨论的扩展需求，本文进一步调研了相关的开放源码系统。为实现特定功能所选用的开放源码系统如表1所示。

表1 扩展技术与解决方案

	目标	解决方案
媒体存档	分布式海量存储	SRB
永久唯一标识符	数字对象的永久命名与解析服务	CNRI Handle System, ARK
选择文件保存格式	适合将来格式的发展	FDA数字存档推荐格式
文件格式注册	文件格式的永久识别	PRONOM, GDFR
格式识别与校验	数据可读、完整性验证	JHOVE
元数据封装	结构化元数据	METS
中文全文索引	全文检索的实现	Lucene其他语言的分词
开放链接	指定开放链接系统的服务	Ananda
认证授权	单点登录	Shibboleth
保存策略	格式有效、媒体有效	多重备份、媒体迁移和翻新、环境封装、按需迁移
资源评价		Log4j日志分析

其中，SRB(Storage Resource Broker，存储资源代理)是美国圣地亚哥超级计算中心研发的数据网格软件，它为用户提供了一个访问文件系统、档案系统、数据库系统等多种异构存储系统的统一接口，屏蔽了存储系统的异构特性。它还支持广域网络环境下多种数据源的访问，提供了复制、复制数据的访问、文件的汇集、分布文件的逻辑集合等功能。

Handle System是由美国CNRI(Corporation for National Research Initiatives)提出的基于因特网的分布式数字对象命名与标识系统，用于在因特网上提供有效的、可扩展的、安全的全球命名和解析服务，它实现了分布式存储数字对象的名称或handle，将handle解析成用于查找、存取和利用资源的有用信息。

ARK是John Kunze受美国国家医学图书馆委托提出的一个开放的、注重实效的、低费用的资源永久性标识解决方案，它建立在URL, handle, DOI, OpenURL等成果之上，综合吸收各家之长克服它们的不足，并对各类资源提供广泛支持。ARK最成功的应用就是加利福尼亚数字图书馆，目前已经分配了80000个ARK。

FDA数字存档推荐格式是美国图书馆自动化佛罗里达中心在其数字保存项目和DAITSS系统(Dark Archive In The Sunshine State)的实施过程中不断修正的适合保存的文件格式。

PRONOM, GDFR(Global Digital Format Registry)都是文件格式注册系统。PRONOM项目是一个有关数据文件格式和支持软件产品的在线信息系统，已包括546种文件格式及其描述和支持的格式工具。GDFR尝试通过注册和维护样本信息，并且提出了新的用户查看和管理信息的想法，以期望最终的注册系统覆盖面广，表达详尽，有严格的有效性，可供公众查询，并且足够支撑我们这个需要存档的年代。

JHOVE(JSTOR/Harvard Object Validation Environment)是JSTOR和哈佛大学图书馆合作研究的格式认证系统，能够支持二进制/ASCII/UTF-8编码的文本格式、GIF/JPEG2000/JPEG/TIFF图像格式、AIFF/WAVE音频格式、PDF、HTML和XML等开放格式，它通过分析文件来自动识别和检验文件格式。

METS由美国数字图书馆联盟开发，它将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生成一个XML文档，用于支持数字对象的存储、传输、转换等操作。它可以标识数字对象内容及其结构，标识元数据并建立与内容实体之间的链接，标识数字对象行为方法，包装二进制的内容数据，因而简单且有良好的开放性、结构性和扩展性。

Lucene是一个基于java的索引工具，可以嵌入到各种应用中实现全文索引检索功能。它完全支持英文分词。如果要在其他语言环境中实现全文索引，还需要通过它的API扩展其他语言的分词器，比如必须要对中文数字资源的索引建立一个中文分词器。

Ananda开放链接系统是由中国科学院国家科学数字图书馆与北京中科软件有限公司共同开发面向中国科学院全院用户的“透明”链接服务。它集成了全文链接、文摘链接、馆际互借和原文传递链接、参考咨询链接以及Google链接等服务。

Shibboleth是一个单点登录的用户身份管理系统，它通过联合管理，由代理器管理多项服务，如认

证、授权、资源发现等,有效地实现内部信息交互。它采用基于属性的访问控制,强化了内部的信任机制,特别适用于联盟和虚拟机构。

Log4j 是 Apache 的一个开放源代码项目,是一个可重用的日志操作类。它由 3 个重要的组件构成:日志信息的优先级、输出目的地、输出格式。优先级从高到低有 ERROR,WARN,INFO,DEBUG,分别用来指定这条日志信息的重要程度;输出目的地指定了日志将打印到控制台还是文件中;而输出格式则控制了日志信息的显示内容[7]。

利用这些开放源码的系统,可以实现一个满足图 2 所描述的技术体系的数字保存系统,其具体实现框架如图 4 所示。

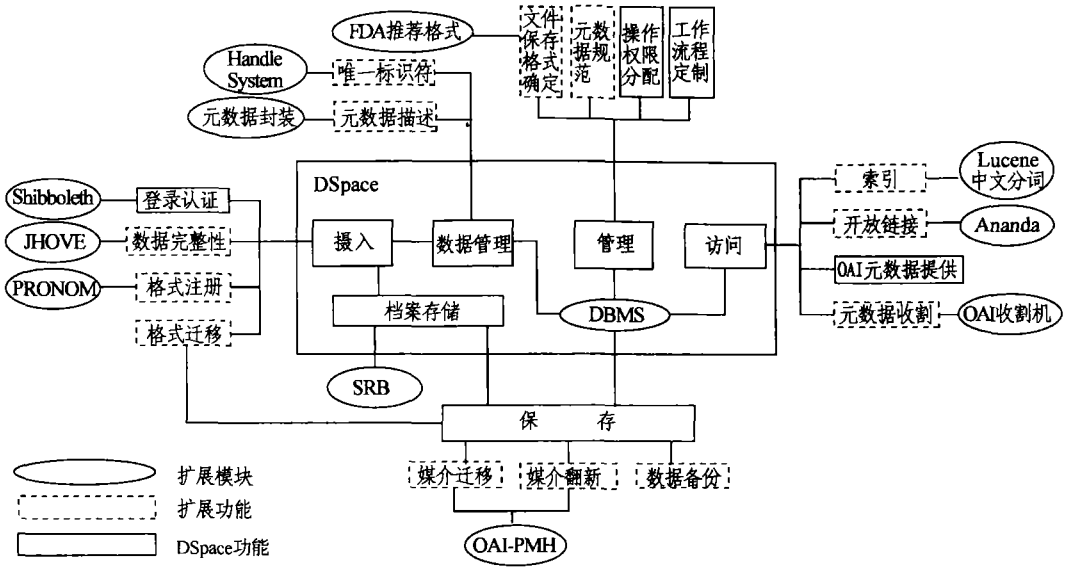

图 4　数字保存系统设计框架

4　总结

本文从降低开发维护成本、缩短开发周期的角度提出了一种基于开放源码的数字保存系统的设计,并提出了相应的技术解决方案。笔者已经在这个系统框架的基础上逐步开发基于 DSpace 的数字保存原型系统,初步实现了 Handle 标识符、RSS、保存文件格式、Lucene 中文分词等技术应用。开发过程表明,利用开放源码系统是高效实现数字保存系统的有效途径。笔者也将密切关注数字图书馆领域开放源码的研究和进展,以进一步充实、完善和提高本文所提出的数字保存系统。

参考文献

1　JISC Circular. Digital Repositories Summary Letter. 2005-03-21. [2006-07-10]. http://www.jisc.ac.uk/uploaded_documents/03-05SummaryLetterFinal.doc

2　Jeff Rothenberg. An Experiment in Using Emulation to preserve Digital Publications. Den Haag . Koninklijke Bibliotheek, 2000. [2006-07-10]. http://www.kb.nl/coop/nedlib/results/NEDLIBemulation.pdf

3　Logging Services-Log4j. [2006-07-10]. http://libraries.mit.edu/dspace-mit/technology/architecture.pdf

4　Apache Lucene. [2006-07-10]. http://lucene.apache.org/java/docs/

5　Migration on Request. [2006-07-10]. http://www.si.umich.edu/CAMILEON/reports/mor/index.html

6　Ir. Raymond Lorie. The UVC: a Method of Preserving Digital Documents - Proof of Concept. 2002-11. [2006-07-10]. http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf

7　Logging Services-Log4j. [2006-07-10]. http://libraries.mit.edu/dspace-mit/technology/architecture.pdf

林　颖　硕士,北京师范大学图书馆工作。通信地址:北京。邮编 100875。

张智雄　博士,研究馆员,中国科学院文献情报中心信息技术部主任。通信地址:北京。邮编 100080。

(来稿时间:2006-07-13)