●余传明 董 慧

开放式存档信息系统及其在数字图书馆的应用*

摘 要 一个开放式存档信息系统是一个包括人员组织、系统、存档数据的整体,它的责任是保存信息并且为指定的团体提供服务。其参考模型旨在建立对数字信息的长期保存和访问,它提供了对存档概念的理解框架和对存档文件的操作,对于制订数字图书馆的元数据标准和功能标准都有重要意义。图 3。参考文献 3。

关键词 开放式存档信息系统 信息模型 功能模型 数字图书馆
分类号 G250.76

ABSTRACT An OAIS system is an unity including human organization, systems and archival data, with the responsibility to preserve information and provide services to specific organizations. Its reference model is to establish long-term preservation and access of digital information, and provide frameworks for the understanding of archival concepts and operations on archival documents. It has great importance to the drafting of metadata standards and function standards for digital libraries. 3 figs. 3 refs.

KEY WORDS OAIS. Information model. Function model. Digital library.
CLASS NUMBER G250.76

1 开放式存档信息系统的诞生

随着数字图书馆研究与应用的深入,数字文献在图书馆及其他文献机构的信息资源中所占比重越来越大。数字文献作为一种新型文献,它与传统文献有着明显不同。数字文献的存储和利用因受硬件、软件与服务系统构成的制约,表现出以下几个方面的特点:

存储数字信息使用的载体一般为磁带和 CD-ROM 光盘。磁带本身就不是永久性存储介质,消磁化对磁带数据的危害比酸性对纸张上数据的危害更大;而虽然 CD-ROM 能使信息在较长时间里得以存取,但是,一旦它的标准被改变,以原有格式保存的信息就存在丢失的危险。

数字信息的阅读与理解需要设备和软件,而由于设备、软件在不断更新换代,所以许多信息最终可能无法读出。

数字信息本身的存储格式及使用技术在不断改变,这也使数字信息的使用寿命受到限制和威胁。

上述原因都带来同样一个问题:数字信息的长期存取与传统文献的保存相比更加困难。如何使数字信息能够在纵向上也具有开放性,使现在的数字信息在未来仍然可以利用,成为摆在数字图书馆研究者面前的一道难题。

在这种情况下,诞生了开放式存档信息系统及其参考模型。开放式存档信息系统(OAIS, Open Archive Information System)是美国空间数据系统咨询委员会(CCSDS, the Consultative Committee for Space Data Systems)提出的一个 ISO 参考模型的草案。根据这个草案标准,一个开放式存档信息系统是一个包括人员组织、系统、存档数据的整体,它的责任是保存信息并且为指定的团体提供服务。这个参考模型旨在建立对数字信息的长期保存和访问,它提供了对存档概念的理解框架和对存档文件的操作。

2 开放式存档信息系统的组成

OAIS 的环境如图 1 所示,它包括 4 个实体:消费者实体、顾客实体、管理实体和存档信息实体。生产者实体提供存档信息;消费者实体利用存档信息,比较典型的一类消费者称为 Designated Community,指代那些能够理解存档信息的消费者;管理实体则负

* 本文属国家社会科学基金项目"数字图书馆相关关键技术研究"(批准号:00BTQ004)的成果。

责决定哪些信息可以存档等。

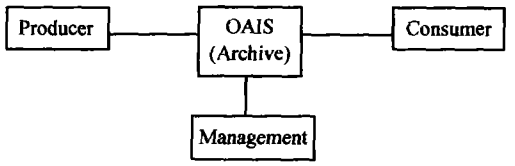

图1 OAIS的环境

2.1 开放式存档信息系统的信息模型

典型的OAIS类型存档模型如图2所示,信息存在于两种形式:物理对象形式和数字对象形式。物理对象形式,包括纸张、图片等;数字对象形式,包括PDF、TIFF等数字形式。两者统称为数据对象(Data Object)。

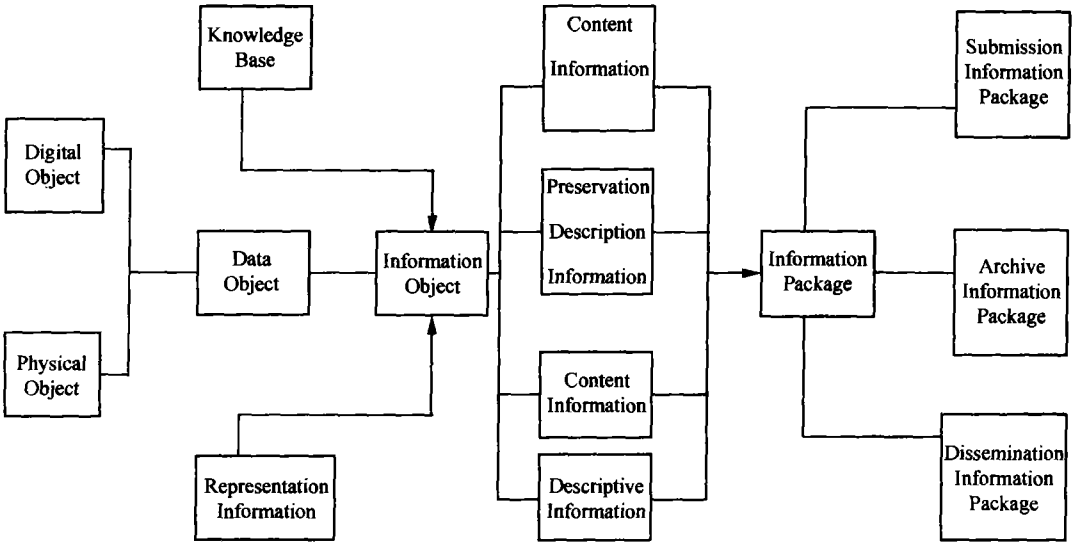

图2 OAIS类型存档模型

消费者实体(包括Designated Community)对于数据对象的理解依赖于两个方面。首先是消费者本身用于理解和阐释信息的知识背景(Knowledge Base)。例如如果这些消费者是JAVA程序员,他们就能够预期理解JAVA代码。但是,消费者实体不可能拥有理解全部数据对象的背景知识,在这种情况下,指示信息(Representation Information)则可以用做补充。例如,如果消费者实体是C程序员,包括JAVA语言句法和规范的背景知识就能够帮助他们理解JAVA代码。背景知识和指示信息结合在一起,代表了消费者实体所能够理解的有用信息,统称为信息对象(Information Object)。显然,信息对象是否有意义仍然依赖于消费者实体本身。

信息对象经过加工后,转化为信息包(Information Package)。信息包中包含4种类型的信息对象:内容信息(CI, Content Information),保存描述信息(PDI, Preservation Description Information),包信息(PI, Packaging Information)和描述性信息(DI,Descriptive Information)。CI指主要的信息内容:数据对象和与其紧密相连的指示信息。PDI指与CI相关的保存信息,包括起源信息、CI的惟一标记符以及CI的有效认证(如检查和数字签名)等。PI是指能够使信息包被识别的一些构件,如光盘的卷和文件的结构等。DI则提供使信息包得以被搜寻和检索的途径。

在OAIS模型中,存在3种形式的信息包:提交信息包(SIP, Submission Information Package),它由生产者实体传递给存档实体;存档包信息包(AIP, Archive Information Package),指存档实体所存储的信息包;传播信息包(DIP, Dissemination Information Package),指根据消费者实体的需求而提供的信息包。

2.2 开放式存档信息系统的功能模型

OAIS的功能模型如图3所示,它包括5个模块:摄取模块(Ingest)、存档模块(Archival Storage)、数据管理模块(Data Management)、检索传递模块(Access)和系统管理模块(Administration)。

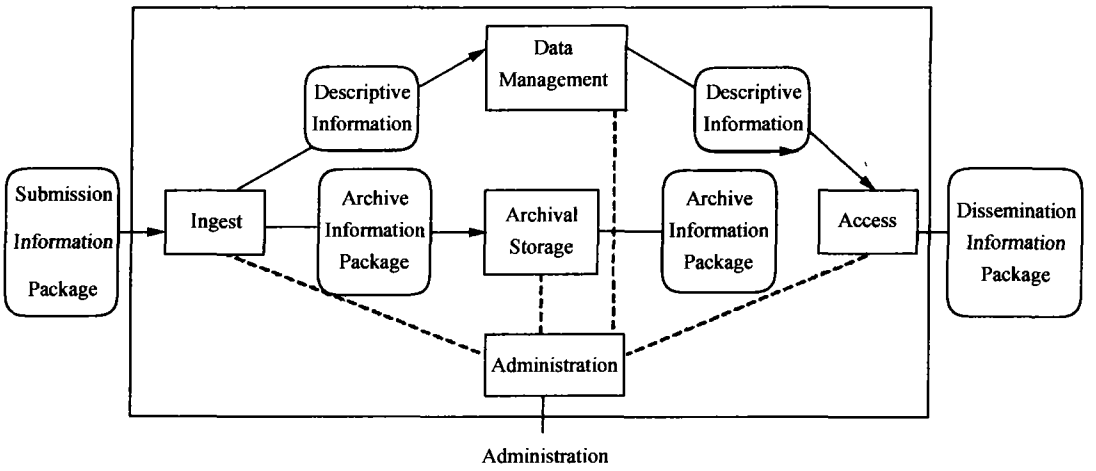

图 3　OAIS 的功能模型

摄取模块从出版商或其他信息提供者处收集或接收按照一定格式组织的提交信息包。这些信息包经过检验后,建立相应元数据,元数据交给数据管理模块,信息包被转换为按照长期保护规定格式组织、包含专门的长期保护处理数据的存档信息包,然后提交给存档存储模块。

存档模块,存储按照 AIP 要求组织的数字信息,包括数据更新、技术仿真或数据迁移,以及建立具体存储与存取系统(如梯次存储系统),并在检索传递模块要求时将 AIP 提供给该模块。在技术仿真或数据迁移过程中,可能形成新的数字内容单元,可能需要与摄取模块协作重新建立有关元数据并送交数据管理模块。

数据管理模块,存储关于数字信息单元的元数据和关于长期保护处理政策、程序、技术和系统的元数据,并提供对这些元数据的基础检索与管理。

检索传递模块,提供用户检索元数据和索取数字信息单元的界面,提供检索机制,并将 AIP 转换为适合用户利用的传播信息包,还可能承担身份认证和授权管理等。

系统管理模块,通过有关规范、程序、工作流等来监测和控制整个系统以及各个模块的运行。

3　开放式存档信息系统在数字图书馆的应用

OAIS参考模型的提出,对于数字图书馆的研究和开发具有十分重要的意义。如上所述,在 OAIS 参考模型中,信息包由 4 种信息对象(CI、PDI、PI 和 DI)组成。根据元数据的定义,这些显然属于元数据的范畴。OAIS 中的信息模型,对于制定适合中国数字图书馆发展的元数据方案具有实际意义。另外,OAIS 的功能模型对于指导和明确数字图书馆的功能具有重要指导作用。

3.1　用于参照制定数字图书馆的元数据方案

在过去的实践中,人们对元数据的研究主要集中在资源的发现上。OAIS 参考模型对此有了突破,它把研究重点放在了资源的长期保存方面。目前,国外对于 OAIS 参考模型中元数据的借鉴已经非常普遍,NEDLIB、CEDARS 以及 PANDORA 等项目,都参照了 OAIS 元数据模型。

以 CEDARS 为例,它是由里兹大学、牛津大学和剑桥大学实施,致力于确认数字信息收藏体系长期保护的战略框架和具体方法,并在此基础上建立支持数字资源长期保存的 CEDARS 元数据,支持数字信息内容的语义提取,还支持有关的描述性、管理性、技术性以及合法性元数据描述。在 CEDARS 中,信息包分为 PDI(Preservation Description Information)和 CI(Content Information)。这与 OAIS 参照模型中的定义完全相同。其中,PDI 又包括确认信息(Reference),例如数字资源标识符及元数据目录;环境信息(Context),描述数据对象与其所处环境和信息系统的关系,也包括与其他信息对象的关系;起源信息(Provenance),描述了数字资源的来源、初始形态等;固化信息(Fixity),描述用以确认信息内容完整性和可信性的信息,例如计算封包内容值、数字签名等。CI 又包括表示信息(Representation),描述存档数字资源的结构和语义信息;数据对象(Data Object),保

存原始数字资源的比特流。

3.2 用于界定数字图书馆的功能

对于数字图书馆功能,国内外研究机构给出了不同的界定。国内比较典型的一种说法是数字图书馆的功能包括5项:(1)各种载体数字化;(2)数据的存储和管理;(3)组织对数据的有效访问和查询;(4)数字化资料在网上发布和传送;(5)系统管理和版权保护。

OAIS的功能参照模型被用于指导确定建立数字图书馆的功能模块。比较典型的项目是欧洲国家版本图书馆(NEDLIB, Networked European Deposit Library),由欧洲7个国家图书馆(荷兰、法国、挪威、德国、葡萄牙、瑞士、意大利)和3家主要出版社(Kluwer, Elsevier和Springer Verlag)参加,其内容包括建立欧洲版本图书馆网络的基础结构,保证电子出版物的长期保存和利用。他们研制的DLS(Digital Library System)系统借鉴了OAIS的参考模型,包括编目、信息采集、DSEP、读者检索权限控制、信息服务等11个模块。其中,DSEP(Deposit System Electronic Publication)负责存储处理和保存功能。DSEP以OAIS作为框架,但它的获取、数据管理和检索功能比OAIS的相应功能范围小,其中有些功能被分到了其他模块里。DSEP通过输入和输出接口来和外界保持联系。输入和输出接口负责将外界接收来的数据格式转换成DSEP规定的格式,并根据用户需求将DSEP的内部格式转换为读者需要的格式。

4 结束语

综上所述,OAIS参考模型对于制定数字图书馆的元数据标准和功能标准都具有十分重要的意义,一旦元数据标准和功能标准得以确定和执行,数字图书馆将不仅在横向上而且在纵向上都具有开放性,数字信息的长期保存和利用问题也得以解决。

值得一提的是,国内对于OAIS的研究目前仍然很少。在具体的数字图书馆实践中,对于OAIS的借鉴将会带来什么样的实际问题,仍然有待于进一步研究。

参考文献

1 张晓林.数字信息的长期保护问题.图书馆,2001(5)
2 杨宗英,郑巧英.数字图书馆研究.大学图书馆学报,2000(1)
3 刘嘉.元数据:理念与应用.中国图书馆学报,2001(5)

余传明 武汉大学信息管理学院情报学2002级博士生。通讯地址:武汉。邮编430072。

董 慧 武汉大学信息管理学院教授,博士生导师。通讯地址同上。

(来稿时间:2003-05-12)

(上接第44页)VISION下一阶段的开发工作将扩展实验数据集。目前VISION中仅仅集成了计算机领域的5000余条书目数据,新词的提取和定位也是在这些书目数据上完成的。我们希望能在更多的领域、规模更大的其它元数据类型上进行实验,最终将VISION推向实用。

4 结语

集成分类法、主题词表和语义元数据构造DL的知识组织系统,为DL提供一个现实可行的知识组织模型,为DL从信息管理向知识管理的过渡提供技术基础。它为当前我国DL业已累积的信息资源提供了基于内容、面向知识的利用和服务手段。VISION原型系统的成果充分说明分类法、主题词表等传统知识组织工具在网络信息环境下仍然有着重要的价值,为了适应数字化、网络化的信息环境,传统图书馆的理论和方法需要不断进行变革与发展。正如国际著名信息学家奈斯比特(J. Naisbit)所指出的:"我们正受信息淹没,但却渴求知识。"DLKOS将为人们提供一个驾驭海量的、日益增长的网络信息资源的知识框架,为解决信息爆炸和信息污染的问题作出图书馆学领域的贡献。

参考文献

1 张晓林.走向知识服务.中国图书馆学报,2000(5)
2 Wang Jun. A Knowledge Network Constructed by Classification, Thesaurus and Semantic Metadata in digital libraries. ASIST Bulletin, 29(2), 2003
3 王军.VISION:集成分类法、主题词表和语义元数据的概念网络.情报学报,2003(4)
4 Sproat, R., and Shih, C. L. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Languages, 4, 4(1990)

王 军 北京大学信息管理系副教授,计算机科学博士。研究方向为数字图书馆、知识组织、信息检索。通讯地址:北京大学信息管理系。邮编100871。

(来稿时间:2003-11-19)