复合数字对象标准中包标准研究*

马翠翠 马建玲

【摘 要】介绍了几种复合数字标准中关于包标准的描述，并指出其在应用领域、目的、兼容性和可扩展性等方面存在的差异。

【关键词】复合数字对象 包标准 OAIS IMS-CP OPC OPF

Abstract: The paper introduces several package technologies in standards of compound digital object, and analyzes their difference in application, object, compatibility and expansibility.

Key words: compound digital object package standard OAIS IMS-CP OPC OPF

复合数字对象是一种复合型的数字对象，即包括文本、图像、音频、视频等类型数字对象的复合体，是各类型数字对象组合起来形成逻辑整体的多个信息单元的集合。比如：由多个章节组成的图书；一个由多种格式和不同清晰度的图片组合成的图片数据对象；由文本及数据集、软件工具、实验、声音记录集合形成的学术出版物[1]。

在复合数字对象中，包含多种格式的文件（如PDF文档或Word文档、JPEG图像或TIF图像）和多种类型的文件（如文本、图像、多媒体文件等），这些文件在网上的存储位置（如存储在数字仓储中、文件系统等）各异、组件之间的关系（比如线性关系、版本关系、引申关系、整部关系等）多样，如何描述这些资源类型之间的关系，将这些分布式、异构的资源形成一个逻辑整体，使之能够被复用、被长期保存，是复合数字对象领域关注的问题之一。包（Package）是解决这一问题的有效途径。本文对IMS-CP、OOXML和OPF等几种标准中的包标准进行了分析比较，希望能够对国内复合数字资源的包标准研究提供参考和借鉴。

1 OAIS[2]

为了更好地理解包概念、包的组成结构，先介绍一下OAIS（Open Archives Information Systems，开放档案信息系统，以下简称OAIS）参考模型。

1.1 类型

OAIS参考模型引入了一个信息包（Information Package，简称IP）的概念，在信息包中存储着数字对象（Information Objects，简称IOs）。OAIS将信息包分为三种类型：

（1）提交信息包（Submission Information Package，简称SIP）：由生产者提交的信息包。

（2）存档信息包（Archival Information Package，简称AIP）：OAIS所存储的信息包。

（3）分发信息包（Dissemination Information Package，简称DIP）：分发给其他用户的信息包。

SIP由创作者创造出来后，被提交存档，存档的SIP被转换为易于存储的AIP，根据用户的请求，OAIS以DIP的方式向用户提供一个AIP的所有或者部分内容，DPI格式的选择往往是适应未来互操作需求的结果。具体流程如图1所示。

信息包在本质上就是各个独立部分（信息对象）所构成的逻辑整体。信息对象包括两方面的内容：数据对象（如MIME优化文件）及其描述信息（XML清单文件）。数据对象和描述清单通常是在一个ZIP包内，但有些情况下则是彼此独立的。图2是根据OAIS参考模型构造的理想信息包的通用结构。

1.2 内容

（1）环境（Environment）：指信息包所存在的环境，它可能是数据库、文件系统或Web服务器。

（2）包（Package）：是一个压缩文件，它由清单文件和其他所有的文件共同构成，包是可选的，文件可以存

* 本文为中国科学院知识创新工程重要方向项目“综合科技资源集合登记系统”研究成果之一。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

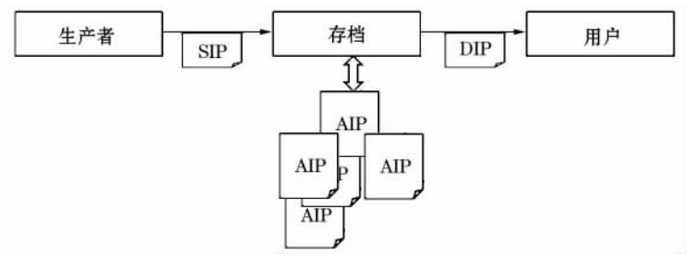

图1 信息包工作流

在于包以外的Web服务器上。

（1）封套（Envelope）：是一个虚拟的组件，当不在包内时，它代表清单文件。

（2）清单文件（Manifest）：描述构成整体的各个部分，包括对象文件信息及其元数据。

①清单文件中的对象文件，是用以描述构成逻辑整体的各个部分，用二进制或ASCII编码。对象文件信息描述文件位置（包内文件和在Web服务器上的外部资源）和文件本身（Base64编码）。

②清单文件中的元数据，用以描述整个文档及各个部分，保护书目信息、出处信息、结构信息和语义信息等。外部引用指向一个Web服务器上的XML文件，内部引用指向包内的文件。

（3）MIME优化文件：是未压缩的比特流，属于某种MIME格式，如JPEG图片。当MIME优化文件不在包内时，将存储于Web服务器上，用URI引用。

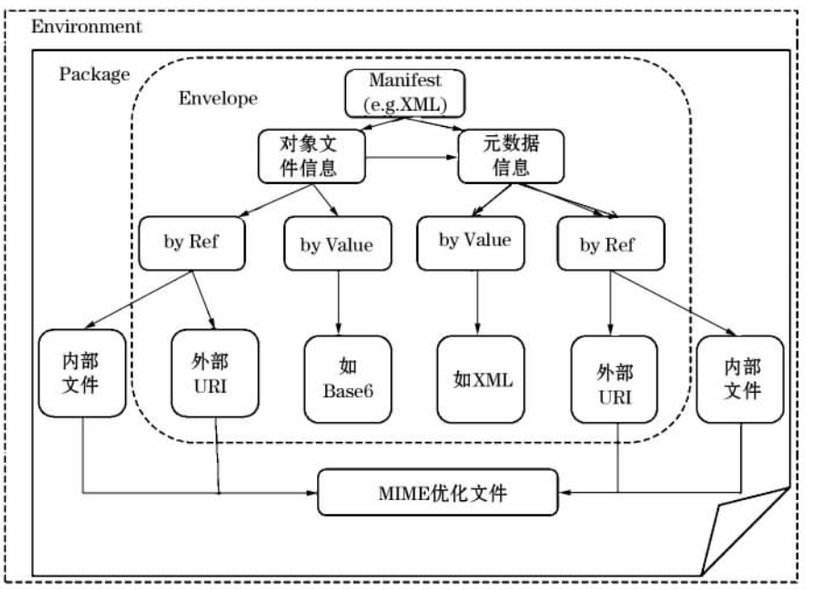

图2 通用包结构

OAIS参考模型中的信息包，基本可以定义本文所研究的包，另外本文所研究的包还具有如下特点：明确边界；描述整体及部分的信息可以出现在包边界之外，具有可扩展性。

2 IMS-CP[3][4]

2.1 IMS-CP概括

IMS内容封装规范（IMS Content Packaging Specification），包含三个文件：IMS内容封装信息模型（IMS Content Packaging Information Model）、IMS内容封装XML绑定规范（IMS Content Packaging XML Binding）和IMS内容封装最佳实践和实施指南（IMS Content Packaging Best Practice and Implementation Guide）。其中，最重要的是IMS内容封装信息模型，是其他两个规范的基础。由于IMS-CP允许E-learning学习内容从一个系统传输到另外一个系统、允许内容在不同学习管理系统中交换，因此它可以作为一个独立的编辑工具。

2.2 IMS-CP内容封装信息模型

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

内容封装信息模型的目的是定义一种能够用来共享和交换学习内容的标准数据结构，从而使学习内容可以在不同的创作工具、学习管理系统和运行环境之间相互交换和使用。一个 IMS 内容包主要包括两部分：清单文件（Manifest），用于专门描述内容结构以及包中资源的 XML 文件，由元数据、组织结构、资源、嵌套的子清单组成；文件资源（File Resources），对应于清单文件中描述的实际的媒体内容、文本文件、图像、其他资源。文件资源可以以子目录的形式组织，如图 3 所示。

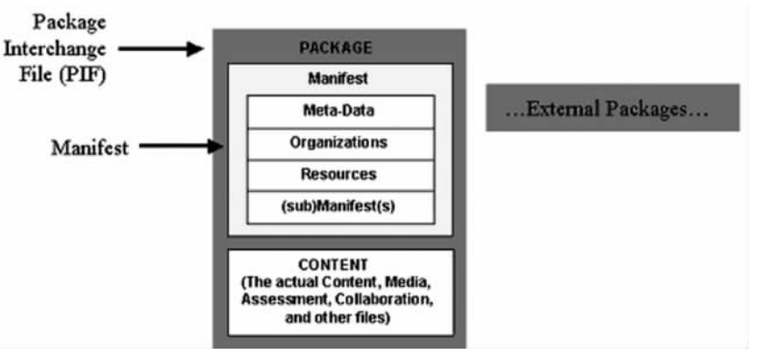

图 3 IMS-CP 的 PIF 结构（程序包交换文件）示意图

2.2.1 内容

（）包：是一个逻辑目录，包含一系列特殊命名的 XML 文件，如 XSD、DTD 文件和实际的文件资源等，这些资源可能在子目录中。每个包都有清晰的、可被解释的边界，这个边界可以是一个 CDROM，也可以是一个符合 RFC1951 的 ZIP 文件。ZIP 文件是可分发的，因此被称为包交换文件（Package Interchange File，以下简称 PIF）。为了让学习对象能在不同平台间传输和交换，包交换文件可以被转换成独立的压缩文件，被其他支持该规范的系统解析和处理。

（）内容即文件资源：包括多媒体资源、文本文件和其他被目录清单所描述的资源。这些资源可被存储在子目录之下。

（）内容清单文件：是包交换文件的重要组成成分，提供了学习对象的描述信息；内容清单文件的结构采用 XML 模式进行描述，每个学习对象的内容清单文件是一个符合该 XML 模式约束的 XML 实例文档。每个清单都包含以下部分：

①元数据部分：对学习内容作整体描述，提供包文件的自检索和自发现的能力。

②组织部分：XML 元素，描述包中物理文件的层次结构和组织顺序。

③资源部分：包含一个清单内所需的所有实际资源和媒体元素，如元数据描述的资源、外部文件的引用等。

④子目录：一个或多个、可选的、逻辑嵌套的清单。

内容清单表现为 XML 文件，包含根元素〈manifest〉。根元素又包含四个子元素：元数据元素〈metadata〉、组织结构元素〈organizations〉、资源元素〈resources〉，以及支持默认显示的 XSLT 样式表元素〈xsl〉。学习对象内容清单的四个模块分别用元数据模式（om.xsd）、组织结构模式（organiztion.xsd）、资源模式（resource.xsd）和样式文件（style.xsl）进行描述。整个内容清单文件编写完成后，连同引用到的物理素材一同封装成包交换文件，便可作为独立的学习单元存入学习对象库。

2.2.2 特点

笔者根据对 IMS-CP 的调研，总结其具有如下特点：

（）来自学习对象领域，由于是非数字图书馆领域，因此尚未引起足够的重视。

（）定义了一个文件能够链接到元数据的嵌入式结构。

（）良好的扩展性，IMS-CP 定义了一个教学测序语言（内容组织），如 IMS 简单测序绑定（IMS Simple Sequencing Binding）；定义了一个 XML 命名空间，组织部分的元素都可以存在它之下，这样对内容包功能进行了有效的扩展。

（）具有独立性，虽然它主要被应用在学习对象领域，但它不依赖于任何教学标准，因此可以被引入到数字图书馆领域。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3 OPC[5][6][7]

3.1 OOXML 基本情况

Office Open XML file format（以下简称 OOXML）最初是由微软开发的一种基于 XML 的文件格式，于 2006 年底成为 ECMA 376（TC45-Office Open XML Formats）标准，并于 2008 年 4 月顺利成为 ISO/IEC DIS 29500：2008，Information Technology-Office Open XML Formats（草案）标准。它是一个文档处理文件，文档中包含的简报、图表等类型的信息将在多种平台上有多种应用。它建立的目的之一是文档的长期保存——确保之前所创立的文件不因技术的进步而无法读取，保证新出现的技术标准能够与原来的文件格式兼容。

OOXML 文档是一个符合 OPC（Open Package Convention，开放数据包规范）的数据包，由一系列部件组成，这些组件中可以包含被嵌入到文档中的图片、音频、视频等资源，并可显示部件之间以及与外部资源的关系，一个完整的 OOXML 文档以 ZIP 压缩包的格式存储。

ISO/IEC DIS 29500：2008 包含四个部分的内容：

（1）ISO/IEC 29500－1：2008 信息技术。文件描述处理语言（Office Open XML 文件格式）的第一部分：基本原理与标记语言参考，这部分定义了一个 XML 词汇表，用来表述文字处理文档、电子表格等。

（2）ISO/IEC 29500－2：2008：信息技术。文件描述处理语言（Office Open XML 文件格式）的第二部分：开放数据包规范（Open Packaging Conventions，以下简称 OPC），它定义了一个通用的文件/组件包机制，其中包含部件之间以及与外部资源的显示关系，一个完整的 OOXML 文档以 ZIP 压缩包的格式存储。

（3）ISO/IEC 29500－3：2008：信息技术。文件描述处理语言（Office Open XML 文件格式）的第三部分：标记兼容性和可扩展性，这部分定义了一个扩展 XML 词汇的通用机制。

（4）ISO/IEC 29500－4：2008 信息技术。文件描述处理语言（Office Open XML 文件格式）的第四部分：过度迁移功能，这部分定义了一系列的 XML 元素和属性，远远超越了过去为微软应用程序服务的 ISO/IEC 29500－1 所定义的内容。

3.2 OPC

OPC 从以下三个方面定义了一个文档的结构，如图 4 所示。

3.2.1 文档结构

（1）内容类型（Content Type）：内容类型部分确定存储在源代码里的文件类型，定义媒体类型、亚型、可选的参数设置。

（2）Relationship（关系）：用以描述资源与目标资源之间连接关系的类型。关系组件直接反映各组件之间的关系而不关注内容部分，所以关系组件是独立于内容部分的，能够快速反应。它包含四个元素：an identifier（标识符）、optional source（可选资源，通常是包或包内的组成部分）、relationship type（关系类型，用 URI 样式定义关系的类型）和 target（目标，指向包内其他组成部分或包外资源的 URI）。

（3）数字结构（Digital Signature）：主要包含验证内容的信息。

3.2.2 数据类型

OPC 包能够存储各种数据类型（文本、图片、XML 等等）的组件。“/_rels”子文件中，扩展名为“.rels”用来存储关系元数据。在 OPC 中，只存在三种名称的文件，即子文件名“/_rels”、文件扩展名“.rels”和文件名“[Content_Types].xml”。

（1）/[Content_Types].xml file：定义了存储在包中的 MIME 媒体类型，根据文件扩展名，/[Content_Types].xml file 定义默认映射。

（2）/_rels：存储包内的各种关系，使之构成一个整体。/_rels 文件中通常包括一个名为“.rels”的文件，“/_rels/.rels”是一个 XML 文件，包级关系（package－level relationships）的起始存于其中。一般情况下，打开一个基于 OPC 的文件，通常从获取“/_rels/.rels”文件开始解读包级关系。

（3）[partname].rels：包内的每个部分都可能与包内的其他部分存在关系，这些关系都存储在 _rels 文件中，因此想查找到某一部分与其他部分之间的关系可以查找 _rels 文件，有关系的文件将以[partname].rels 的名称存在于 _rels 文件中。

3.2.3 特点

OPC 的特点表现在：

（1）鼓励文件进行分块（chunk），这样能够减少文件的损坏，更好地进行数据访问。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

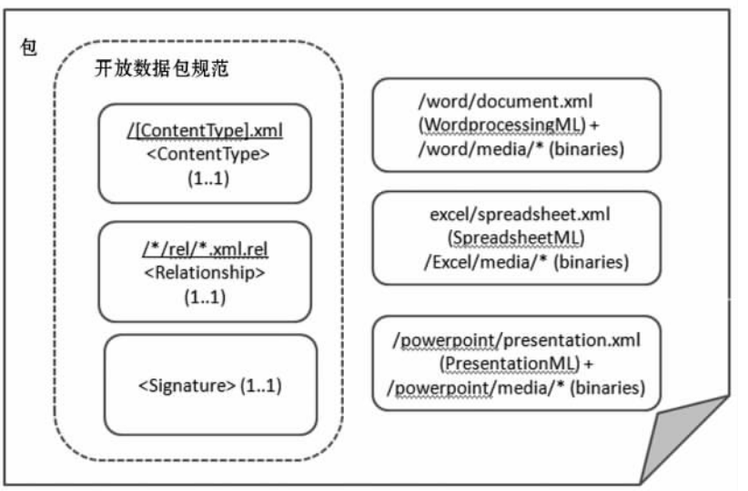

图 4 OPC 包结构

(3) 在包中使用了一个单独的“关系”(Relationship)文件保存对外部文件的超链接方式或链接式的引用，简化了对链接的编辑和管理。

(4) 它在某些方面与现有的 ISO 标准不兼容，某些标准的使用与 W3C 的推荐标准也不兼容。

4 OPF[8]

4.1 Open eBook 概括

Open eBook 格式由国际数字出版论坛(International Digital Publishing Forum，简称 IDPF，是数字出版产业的贸易标准协会)创建并维护。Open eBook 格式由三个部分构成：开放出版结构(the Open Publication Structure，以下简称 OPS)，定义了构造 OPS 内容文档所必需的标记；开放数据包格式(Open Packaging Format，以下简称 OPF)，定义了一种机制，使待出版的著作的各个部件符合 OPS 标准，最后将这些部分打包为 OPS 出版物；开放容器格式(Open Container Format，以下简称 OCF)，定义了一种机制，使 OPS 出版物的各个组件能够形成一个独立个体。

4.2 OPF

OPF 如前所述，它定义了一种机制，使待出版的著作的各个部件符合 OPS 标准，最后将这些部分打包为 OPS 出版物。OPF 定义了一系列元素用以实现这些功能，如图 5 所示。

4.2.1 结构

(1) 包名称(Package Name)：OPS 出版物作为一个整体的标识符，它是 OPF 包的根元素，其他元素是嵌套在它里面的。

(2) 元数据(Metadata)：必选元数据元素能够提供整个出版物整体的各方面的信息。

(3) 文件清单(Manifest)：文件清单中主要包含了一系列文件或者是 item 元素，每个 item 描述一个文件、一张图片、一个样式表及其他任何被认定为是出版物组成部分的组件。item 元素必须包含的属性有 id，href 和 media-type。文件清单中的文件排序是任意的，并没有特殊规定。

(4) 脊(Spine)：该元素定义了出版物的阅读顺序。Spine 中包含一个或多个 itemref 元素，每一个 itemref 元素都与指定的文件清单中 OPS 内容文件相对应。由于这种对应顺序，OPS 文档内容的线性阅读顺序是通过 itemref 的顺序来组织的。

Spine 中有一个 NCX，NCX 是 Navigation Center eXtended 的简称，它的目的是为了使导航更具有便捷性和获取的快捷性，已经被 DAISY 联盟认定为标准并负责实施。NCX 覆盖整个出版物的层次结构，用户通过它可以导航整本出版物。它与目录(able of content)相似，用户可用它迅速定位到出版物的相关章节，但它的内容要远多于目录中所抽离出的章节、图表等信息，它可以被可视化为一个类似于 PC 文件夹树形目录，通过对文件夹的不断展开，发现有用信息并定位。

(5) 导览(Tour)：在这部分，会将用户可能感兴趣的部分抽选出来做一个集合，使用户能够迅捷地发现这些

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

资源。每个 Tour 中都包含一个或多个 site 元素，每个 site 元素必须具备 herf 属性和 title 属性。Tour 已经过时，它并不是包中的必备元素。

（3）指南（Guides）：用来标识书籍结构的各个组成部分，使阅读系统能够准确地定位到相应的部分。通过 Guides 可以将用户指引到文章的具体章节、图表、参考文献等部分，但 Guide 并不是必备元素。

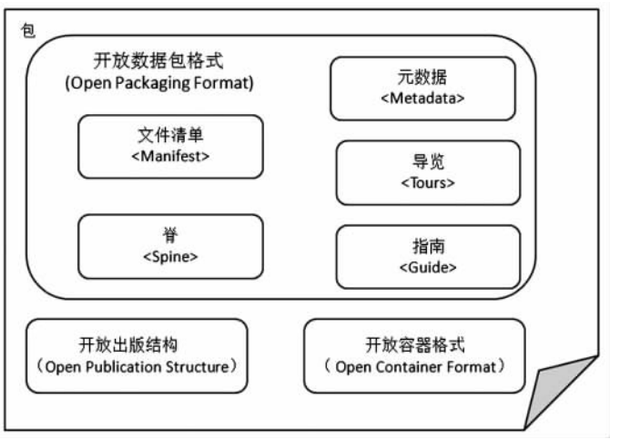

图 5 OPF 包结构

4.2.2 特点

OPF 具有如下特点：

（1）描述了应用 Open eBook 格式出版的电子出版物的各个组件，如标记文件、图片、导航结构等。

（2）提供了出版级的元数据。

（3）指定了出版物的线性阅读顺序。

（4）当环境不支持对 OPS 扩展时则提供后备信息供使用。

（5）通过 NCX 提供了指定内容声明表机制。

5 IMS-CP、OPC 与的 OPF 比较

5.1 共同点

IMS-CP、OPC、OPF 三者尽管研发的领域不同，但它们都属于复合数字对象包标准。首先，它们都能够将不同类型的数字对象封装在一个 ZIP 包内，实现内容的传递、识别、解析和共享；其次，这些标准都是基于 XML 的，所以可以用基于 XML 的工具进行解析、验证、编辑和转换等操作；再次，这三个标准都是使用 URI 作为唯一标识符；最后，这三者都不能进行包外存档。

5.2 不同点

（1）发端于不同的领域：IMS-CP 发展于分布式教育资源领域；OPC 始于文档对象领域；OPF 出现于数字出版领域。

（2）目的不同：IMS-CP 的目的是为了将这些分布的教育资源作为一个整体提供给网络学习用户使用；OPC 是为了解决由于应用不同办公软件导致文档间缺乏互操作性和可交换性的问题；OPF 目的是为了数字出版物在不同的阅读系统中都能够得以准确地读取。

（3）从兼容性上来看，IMS-CP 与 OPF 是不具有兼容性的，而 OPC 与现有的 Microsoft Office 文档向后兼容。

（4）从扩展性上看，IMS-CP 和 OPC 都具有良好的扩展性，都是通过对外部建构的扩展来实现的，而 OPF 则不具有扩展性。

6 结语

复合数字对象是目前数字对象的主要存在形式，如何实现复合数字对象的组织、共享、重用一直是该领域关注的问题。目前关于这方面的研究已经取得了一定的进展，但关于这些技术之间的扩展性研究、比较性研究还相对较少，还需要进一步加强。

（下转第 97 页）

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（5）信息分析工具。面向科研团组开展的信息服务工作不仅包括信息资源的提供，更要重视信息分析服务。因此，信息服务工具必不可少，比如利用 ISI Web of Knowledge 数据库、incites 等文献分析工具对学科理论进展和科研团组的论文产出情况进行分析；利用德温特专利数据库对专利技术创造情况进行分析；利用 SPSS 对相关信息进行数据分析等。

4 开展科研团组信息服务的意义

科研团组从事的是一种知识创造工作，对信息具有强烈的需求。图书馆作为信息服务部门，加强对科研团组的信息服务，不仅为科研团组的发展提供了必要的资源保障，也是图书馆工作的一大进步，具有重要的意义[6]。首先，拓展了图书馆的服务对象，丰富了图书馆信息服务的内涵，并使图书馆的信息资源得到更充分的开发和利用；其次，提高了图书馆的情报分析能力。科研团组需要的信息资源是经过加工与提炼的科技情报，因此对科研团组的信息服务工作必将促进图书馆情报分析能力的提高[9]；再次，强化了图书馆知识服务特色。知识化服务是图书馆信息服务的总体发展趋势，面向科研团组的服务以科研价值较高的科研知识为主，因此强化了图书馆的知识服务特色。最后，提高了图书馆的社会价值。图书馆通过对科研团组开展的信息服务，间接参与了国家的科技发展战略，为科学研究的发展提供了资源保障，这使图书馆的社会价值得到了进一步提高。

注释

[1]裴长洪，王镭．试论国际竞争力的理论概念与分析方法．中国工业经济，2002（4）：41－45

[2]谢海波．科研团队的问题与对策探析．辽宁行政学院学报，2010（4）：70－71

[3]“长江学者和创新团队发展计划”创新团队支持办法．http：//baike.baidu.com/view/3443799.htm

[4]蒋日富，霍国庆，谭红军，郭传杰．科研团队知识创新绩效影响要素研究——基于我国国立科研机构的调查分析．科学学研究，2007（2）：364－372

[5]李冬琴，李靖华，吴晓波．我国高校和科研机构科技竞争力的比较分析．科学学研究，2003（4）：378－384

[6]邵荣，李大钧，李威，张军．面向课题组的学科化服务模式探讨．图书情报工作，2009（2）：86－89

[7]朱强，孙卫，赵亮，马瑞，洪光宗，孙一钢．以开放的心态迎接新的信息技术．中国图书馆学报，2010（5）：77－94

[8]王涛．影响图书馆未来发展的新技术．今日科苑，2010（22）：204－204

[9]朱建贞．图书馆服务理念的新发展——谈知识经济下图书馆的知识服务．科技情报开发与经济，2008（17）：45－46

[10]朱建贞．高校图书馆科研定题服务工作理论与实践探讨．科技情报开发与经济，2008（17）：45－46

刘的帝　中国科学院国家科学图书馆，硕士研究生。

杨志萍　中国科学院国家科学图书馆成都分馆，副馆长，硕士生导师。

（止接第 56 页）

注释

[1]马建霞．数字仓储中复合数字对象相关标准比较研究．现代图书情报技术，2009（4）：33－39

[2]Reference Model for an Open Archival Information System（OAIS）．http：//www.google.com.hk/url？q＝http：//public.ccsds.org/publications/archive/650x0b1.PDF＆sa＝U＆ei＝MF1zTtuPPIW0iQehIbSxDQ＆ved＝0CBAQFjAA＆usg＝AFQjCNEg2TT1gOm－XJI－HdMbg6hmPApJTQ，2011－09－02

[3]IMS-GLC：Conten Packaging Specification．http：//www.imsglobal.org/content/packaging/，2011－09－06

[4]IMS Content packaging．http：//edutechwiki.unige.ch/en/IMS_Content_Packaging，2011－08－27

[5]Ecma international．TC45－Office Open XML Formats．http：//www.ecma－international.org/publications/standards/Ecma－376.htm，2011－09－09

[6]Walter Ditch XML-based Office Document Standards．http：//www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanning/hs0702.aspx，2011－09－12

[7]Open Packaging Conventions－Wikipedia，the free encyclopedia．http：//en.wikipedia.org/wiki/Open_Packaging_Conventions，2011－09－11

[8]Open Packaging Format（OPF）2.0v1.0．http：//old.idpf.org/2007/opf/OPF_2.0_final_spec.html，2011－09－12

马翠翠　中国科学院国家科学图书馆兰州分馆。

马建玲　中国科学院研究生院。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net