

数字资源共享平台框架的设计和实现^{*}

朱义华，郭昭辉，武港山

(南京大学 计算机科学与技术系，江苏 南京 210093)

摘 要：从数字资源共享平台的需求入手，分析了目前建设数字资源共享平台面临的一些问题，提出了一种数字资源共享平台的框架设计，并基于 CNRI Handle System 和 OAI-PMH 给出了该框架的技术实现方案。最后介绍了一个实现实例——中国大学数字博物馆共享平台。

关键词：资源管理；互操作；共享平台；句柄系统；开放文献预研元数据获取协议

中图分类号：TP311 文献标志码：A 文章编号：1001-3695(2007)05-0214-04

Design and Implementation of Sharing Platform of Digital Resource

ZHU Yi-hua, GUO Zhao-hui, WU Gang-shan

(Dept. of Computer Science & Technology, Nanjing University, Nanjing Jiangsu 210093, China)

Abstract: Author analyzed the requirements of a SPDR and some challenges may met during the building a SPDR. Then presented a framework of SPDR and gave a design of this framework based on CNRI Handle System and OAI-PMH. At last introduced implementation of this framework named Sharing Platform of Digital Museum of Chinese University.

Key words: resource management; interoperability; sharing platform; handle-system; OAI-PMH

0 引言

随着信息载体数字化和网络化的迅速发展,越来越多的信息资源成为网络上可以被描述和访问的资源,从简单的文本文件到多媒体声音、图像资源,甚至网络服务(Web Service)等都加入了这一行列。为了高效率、高质量地收集、保存、管理这些资源,数字图书馆、数字博物馆等一些新的信息资源组织形式应运而生;同时随着数字化信息资源的快速增长以及各种管理数字化信息资源系统的建设,为了有效地整合并合理配置分散的数字资源,扩大数字资源的覆盖范围和服务范围,数字资源共享平台的建设已经成为信息资源数字化建设的重要任务。从功能需求上看,数字化资源的定位以及不同系统、不同数据拥有者之间如何实现资源的共享和互操作已成为数字资源共享平台建设的重要问题。

目前国外在这方面的工作已有一定的基础,特别是在数字图书馆领域,通过 Z39.50、OAI 等协议集成异构资源系统成为构建数字图书馆资源共享平台的常用方式^[1]。国内在数字图书馆领域一般采用简单导航整合模式(即在 Web 页面上提供按字母或主题分类入口的方式),还有一部分采用构建中间层的模式(如江苏汇文一站式检索系统和清华同方的异构数据库统一检索平台),避免了需要逐个登录数据库、输入检索条件的麻烦。一方面这些技术有着不可避免的缺陷(如中间层技术导致检索效率低下的问题)需要寻找更好的整合模式;另一方面在非数字图书馆领域,各领域系统的建设仍然比较孤立,共享能力还比较低。

1 共享平台的需求分析

数字资源共享平台的基本功能是通过不同系统之间数字资源的互操作,实现数据资源整合、信息发布,并对外提供资源检索服务。具体系统建设时,根据应用需要,数字资源共享平台还会包括用户权限控制、资源版权控制等处理机制。数字资源整合是依据一定的需要,对各个相对独立的数字资源系统中的数据对象、功能结构及其互动关系进行融合、类聚和重组,重新组合为一个新的有机整体,形成一个效能更好、效率更高的新的资源体系^[2];信息发布和检索服务则是在资源整合的基础上,向用户展示资源相关信息,让用户比较方便地使用数字资源。

从最终用户的角度出发,共享平台必须保证系统的透明性,即用户可以在一个系统上浏览整个平台中的数字资源;用户只要一次检索便可以覆盖平台中所有的系统;无须知道资源的位置就能够获得需要的数字资源等。从平台内的系统出发,平台内部的单个系统内部各组件间以及各系统间必须能够交互,以实现数字资源元数据的互操作以及检索服务的互操作。为了便于系统间互操作的实现,资源的发布必须满足一定的发布标准。这样各系统才能获得其他系统发布的资源。同时为了在获得资源后有效地进行整合,那么就需要知道获得资源元数据的语义,必须制定统一的元数据描述标准。为了避免平台内部冲突,保证平台的透明性,方便用户的使用,数字资源的定位机制以及共享平台的服务和管理规范也是必需的。

鉴于上面的需求,笔者认为平台的建设需要分层次、分模

收稿日期：2006-01-06；修返日期：2006-04-12 基金项目：国家“863”计划资助项目(2002AA117010-10)；2005 年教育部资助项目；2005 年科技部科技基础条件平台资助项目(2005DKA64300)

作者简介：朱义华,硕士研究生,主要研究方向为 Web 信息检索、自然语言理解(zhuyihua@graphics.nju.edu.cn)；郭昭辉,硕士研究生,主要研究方向为 Web 信息检索、自然语言理解；武港山,男,教授,博士,主要研究方向为 Web 信息检索、多媒体信息处理。

块来实现不同的机制, 从而保证平台内部各系统既能保持其自主性、独立性, 又能实现系统间的互操作性、共享性。下文提出了一种数字资源共享平台的建设框架, 并主要从资源的发布、定位以及平台必须提供的服务这三个方面分析其技术特色和实现方案。

2 共享平台框架设计

数字资源共享平台总体框架由存储层、描述层、组织层和应用层四个层次组成, 如图 1 所示。

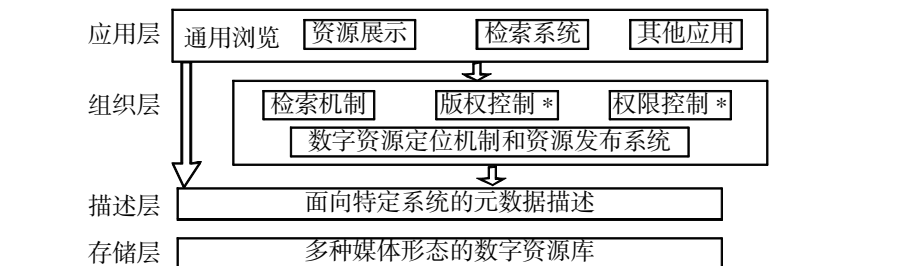


图 1 数字资源共享平台的总体框架(* 表示可选)

存储层中存储的是关于数字资源的多媒体描述信息, 比如文本、图片、视频、3D 模型等。为了便于保存各个媒体类型之间的关系以及其他诸如用户对某个资源拥有的权限等关系, 可以采用关系数据库来存储这些信息。描述层是指按照某特定领域制定的元数据描述标准来描述数字资源; 确定的元数据格式是组织层中资源发布系统以及互操作的基础。组织层中, 首先必须有资源发布系统以及数字资源定位机制。其中资源发布系统一方面面向浏览用户, 便于用户查看数字资源; 另一方面面向共享平台中的其他系统, 便于系统间的互操作。定位机制则是要防止资源命名冲突, 同时提供对名称快速简单的解析功能。基于此, 还应该提供检索机制以方便用户对数字资源的查找(检索机制必须是多系统之间的协同检索, 这样才能保证应用层的检索系统能够为用户提供透明性), 同时还可以提供其他机制, 如用户权限控制、资源版权控制等。应用层则通过友好的用户接口直接面向用户提供了一些必要的服务, 如资源的展示、检索服务等, 应用层通过组织层来隐藏细节, 提供透明性。

共享平台的分层设计以及各层功能的明确, 便于研究者根据自身的建设需求运用相关技术快速地实现系统; 各个模块的相互独立, 也便于研究者根据实际需求进行取舍。目前, 在上述层次的几个方面都分别提出了一些相应的解决标准和方案。数据内部格式方面即元数据描述方面现在有一些比较常用的元数据格式、如 US MARC 格式、都柏林核心数据 (Dublin Core)、VAR 核心类目 (VAR Core Category)、艺术作品著录类目 (Categories for the Description of Works of Art)、REACH 著录单元集合 (REACH Element Set) 等^[3]; 唯一标志符方面有目前正在使用的用以标志期刊和期刊所包含内容的标准化可变长标志符 SICI (Serial Item and Contribution Identifier)、参照 SICI 而提出的用以标志图书及图书具体内容的可变长标志符 BICI (Book Item and Contribution Identifier)、用以标志出版和销售过程中涉及的各种内容单元的 PII (Publisher Item Identifier), 还有包括了永久性命名和解析服务系统的 PURL (Persistent URL)、Handle、DOI (Handle 系统在出版行业的应用) 等^[4]; 在资源发布方面, 为了实现互操作, OAI 和 OpenURL 逐渐被采纳成为协议标准^[5]。

下面从上述的几个方面入手, 结合每个方面比较成熟和优

秀的技术, 提出了上述平台建设框架的技术实现方案。

3 共享平台框架的技术实现方案

上述共享平台框架中定位机制不仅需要能够唯一标志对象, 更需要持久、可操作地标志对象以满足不同系统间的互操作, 所以笔者选择采用了 CNRI Handle System^[8]。数字资源发布方面, 笔者采用了能够适应元数据的复杂性、多样性及灵活性, 对不同资源和不同组织的元数据进行集成并实现互操作的基于开放文献预研 (Open Archives Initiative) 的元数据互操作协议 (OAI-PMH)^[9]。检索服务方面, 笔者通过 OAI-PMH 中的服务提供者提供了检索服务, 又应用 Web Service 技术提供检索服务接口, 从而实现了各个系统间的协同检索。

3.1 数字资源的唯一标志和解析系统——Handle System

Handle System 最初是由美国 CNRI (The Corporation for National Research Initiatives) 提出并实现, 建立在 Internet 架构之上的一个分布式的信息系统, 用来提供有效的、可扩展的、可靠的进行名称解析和管理的全球名字服务^[6]。

Handle System 包括一套开放的协议、名称空间和协议的参考实现模型。Handle System 允许网络上不同的计算机存储数字资源的名称 (或 Handle), 以及将这些 Handle 解析成与数字资源相关的信息片断, 以满足对该数字资源的访问、定位和使用。这些信息片断可以被授权用户根据实际需要修改和维护, 用户在分布环境下管理信息片断, 从而保证名称可以持久地标志对应的数字资源。

Handle System 由解析系统和管理系统两大部分组成。解析系统把用户提供的 Handle 解析成与该 Handle 相关的信息, 以使用户定位、访问和使用数字对象; 管理系统则负责提供 Handle 的用户对 Handle 的有关信息进行编辑和修改的功能, 以维持名字与实际对象之间的关系。

Handle System 的名称空间由两个部分组成, 即命名机构标志符和本地唯一标志符。其名称服务是 GHR (Global Handle Registry, 通用名称登记系统) 和 LHS (Local Handle Service, 本地名称服务系统) 的两层结构。GHR 提供命名机构标志符的解析服务, 目前是单一的服务, 由 CNRI 管理; 它将所有的用户解析请求分发到具体负责的 LHS。LHS 提供它所管理的名称空间下的唯一标志符的解析服务。图 2 是一个名称为 dmcu.nju/9802600040088 的 Handle 解析过程。



图 2 Handle System 的名称解析过程

在数字化信息资源共享平台的建设框架中, 客户端可以通过三种方式解析 Handle: 代理服务器的方式使用 Handle System, 此时浏览器端不需要使用任何额外软件; 直接解析, 此时用户需要安装浏览器的插件; 使用 CNRI 提供的类库, 此时用户可以在专业的客户端中使用解析服务。

3.2 数字资源互操作协议——OAI-PMH

开放文献预研 (OAI) 框架是在 1999 年 10 月一次学术出版联合组织的会议上提出的。OAI 的制定最初是针对学术性电子化预刊本的互操作及检索, 但这与各类型数字化信息资源

管理系统建设中所遇到的元数据互操作问题相类似,所以在 2000 年上半年, OAI 使用范围迅速扩展到这些领域。2001 年 1 月, OAI 发布了名为 Open Archives Initiative Protocol for Metadata Harvesting(OAI 元数据获取协议) 的网络通信协议, 为网络上元数据的互操作问题提供了一种可行的解决方案^[7]。

OAI-PMH 提供了一个基于元数据的独立于具体应用的互操作框架。在此框架中有两个级别的参与者, 即数据提供者(Data Providers) 和服务提供者(Service Providers)。数据提供者主要是存储和管理数字资源, 并遵循 OAI-PMH 协议发布这些资源的元数据, 发布的元数据格式可以多样化, 如 DC、MARC 等, 构成资源发布系统; 服务提供者(Service Providers) 以 OAI-PMH 为基础获取并整合元数据来建立增值服务, 主要是检索服务。

数据提供者拥有信息仓储(Repository)。仓储由多个条目(Item) 组成, 一个条目是对一个具体资源(Resource) 的描述, 它具有一个唯一的标志符。一个条目可由多种元数据格式描述, 每一种元数据描述为一个记录(Record), 但它们均对应同一个标志符。由于每个数据提供方各自拥有自己的元数据, 其元数据的格式也不尽相同, OAI 规定数据提供方可发布多种格式的元数据, 但一定要有一种是以都柏林核心集元数据格式提供元数据。元数据被封装为 XML 格式进行通信, 即数据提供者提供给服务提供者的是 XML 元数据记录。为了提供选择性获取, 数据提供者还提供以日期为基础(Data-based) 或特定集合为基础(Set-based) 的元数据获取方式, 使用者可以比较精确地描述欲获取资料的范围。

服务提供者通过 OAI-PMH 协议的核心——六个命令动词(Verb) 向数据提供者获取资源的元数据。这六个动词分别是: GetRecord, 用于从仓储中获取一条元数据记录; Identify, 用于获取有关仓储的信息; ListIdentifier, 仅返回头部而不是记录本身; ListMetadataFormats, 用于在仓储中检索可以获得的原数据格式; ListRecords, 用于从仓储中获取记录; ListSets, 用于返回仓储的集合结构(Set Structure), 有利于选择性获取。其详细应用可参照“*The Open Archives Initiative Protocol for Metadata Harvesting*”^[9]。

数据提供者与服务提供者之间通过 OAI Request 和 OAI Response 实现消息的传递, Request 由 HTTP 的 GET/POST 方法实现, 每个请求包括一个基本的 URL 地址以及关键字、参数; 数据提供方在接收到数据获取请求后, 以 XML 记录将元数据返回给请求方。其协议工作模式如图 3 所示。

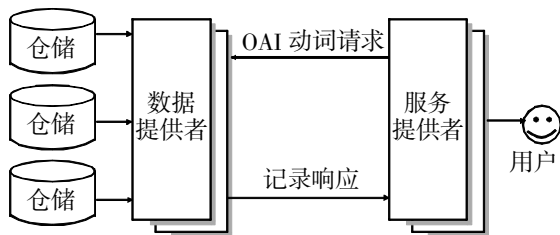


图 3 OAI 元数据互操作框架

目前 OAI-PMH 已经有相对成熟的实现, 在本文基于此框架实现的中国大学数字博物馆共享平台(见 <http://dmcu.nju.edu/cn>) 中, 采用了 OAI-PMH 的一个开源实现 OCLC 的 OAI-Cat^[10], 并作了一部分修改。

3.3 检索服务

在本文提出的这个数字资源共享平台框架中两种检索服

务: OAI 服务提供者提供的检索服务; 独立于 OAI 框架, 由元数据拥有者通过 Web Service 接口提供的元数据检索服务。这两种检索相辅相成, 能够满足不同的需求, 其结构如图 4 所示。

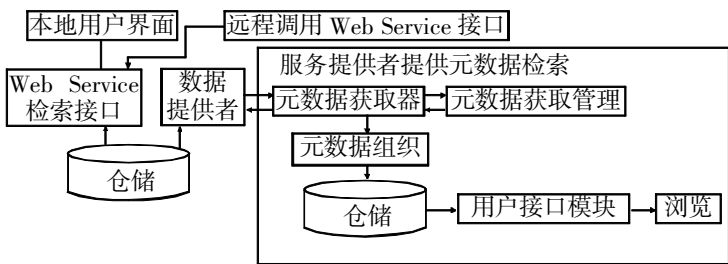


图 4 两种检索服务体系结构图

OAI 服务提供者提供的对元数据的检索服务体现了互操作的价值。服务提供者在本本地建立获取器程序, 执行定期定量的元数据获取任务, 有选择性地收割数据提供者发布的元数据, 经过整合映射之后存放在本地, 并对外提供检索服务。同时, 数据提供者由于某种原因不一定把所有的数字资源都通过 OAI-PMH 发布出来, 或者仅仅发布资源的部分元数据, 即数据服务者无法获得某些资源的元数据或者仅仅获得某些资源的部分元数据而不是其详细的元数据。但数据提供者又希望提供检索服务告知检索用户是否拥护其关心的数字资源, 所以此时可以在数据提供者本地提供检索服务并以 Web Service 方式提供接口。这样既能保护元数据, 又能让用户获得某些信息。

4 实例

2003 年 7 月由教育部科技司主持, 18 所高等院校承担的“现代远程教育网上公共资源建设—大学数字博物馆建设工程”项目通过教育部验收, 初步建成了地球科学、生命科学、人文科学、工程技术四类大学数字博物馆。各博物馆的建设内容以自有馆藏实物为基础, 由相应领域的专家审定或撰写文字材料, 并与大学相关专业的教学科研互为补充, 其科学性和权威性都很强。大学数字博物馆从建成到现在虽然已经取得了一定的社会效益, 但由于时间和其他客观条件的限制, 建成的博物馆系统还存在一些问题: 缺乏统一技术标准规范、共享服务薄弱等。为了进一步实现资源的优化整合, 更大地发挥博物馆的作用, 以及高校学科、人才、资源的综合优势, 教育部开展了“中国大学数字博物馆共享平台与规范标准”项目, 拟在制定相关建设规范与标准的基础上, 进一步实现资源的优化整合, 建设一个统一的数字博物馆共享平台, 从而实现资源共享, 提供综合信息服务。

基于上述框架, 在制定了人文艺术、生命科学、地球科学、科学技术这四个学科领域的数字资源元数据标准的基础上, 加上相关的学术交流与合作提供了快捷、方便的途径, 更加合理、高效地利用了各藏馆资源。数字博物馆藏品有以下两个特点:

藏品名称需要有稳定性, 这样研究人员不必担心藏品资源属性的改变, 只需要记住藏品的名称即可; 每个藏品可以有多种展现方式, 研究人员可以按照自己的需求获得某藏品的某种展现方式。鉴于这样的特点, 笔者按照 Handle System 的标准建立了中国大学数字博物馆共享平台中数字资源的命名系统, 开发了 DMOI(Digital Museum Object Identifier) 服务, 并采用代理服务器的方式解析 Handle。

Handle 协议定义了 Handle 的构成规则。Handle 通常情况下由命名授权 (Naming Authority, NA) 和本地名字 (Local Name) 构成。NA 实际上是一个名称空间, 保证了不同组织的 Handle 的唯一性。一个组织有了 NA 以后, 就可以在这个 NA 下注册 Handle。DMOI 的生成采用了与 Handle 相同的规则。DMOI 也是由两个部分组成, 即 NA 和本地名字。NA 可以具有层次结构, 笔者在 DMOI 系统中规定所有分馆的 NA 都在 dmcu 命名授权下, 如南京大学数字博物馆的 NA 是 dmcu. nju, 复旦大学数字博物馆的 NA 是 dmcu. fudan 等; 本地名字则采用原系统中的藏品 ID, 这样就得到了每个藏品唯一的标志符。当然此平台还没有在 GHR 注册笔者使用的 NA, DMOI 系统目前只是本共享平台内相互协作的一个系统, 与外界并没有关联。

笔者基于 Handle System 构建了 DMOI 服务解决藏品的命名、解析定位问题, 同时基于 OAI-PMH 实现了各个馆之间的用户权限控制和资源的版权保护机制。以及其他的诸如面向藏品的自动回答系统、藏品 3D 数据获取动态展示、虚拟现实等机制。笔者实现了中国大学数字博物馆共享平台 (<http://dmcu.nju.edu/cn>) 。

数字博物馆为很多学科的研究人员提供了丰富且可靠的科研资料, 共享平台的建设更为互操作问题。

OAI 协议本身已经有不少的实现。本系统在实现中采用了开源的 OAICat^[10] 包。由于 OAICat 仅仅支持从一张数据表中获得不同的元数据格式, 而不支持从不同的数据表中获得不同的元数据格式。笔者在该软件包的基础上进行扩展, 实现了从多张数据表中获得不同元数据格式的功能, 并将其应用到本平台中。

中国大学数字博物馆共享平台中存在着两个不同的角色, 即分馆和地区中心。各个分馆发布的数据遵循核心元数据格式(指定为 Dublin Core) 和扩展元数据格式。扩展元数据格式满足该藏品所属类别(人文艺术、生命科学、地球科学、科学技术) 的领域元数据标准。分馆和地区中心都是支持了 OAI-PMH 接口的数据提供者和服务提供者。其中分馆服务基于本馆数据提供, 而地区中心服务基于的数据来自所有分管数据提供者提供的数据。

基于这样的共享平台框架, 笔者满足了大学数字博物馆

共享平台基本服务建设的要求并达到预期的目标, 实现了相关资源的共享, 提供了相应的服务。

5 结束语

基于笔者提出的这样一个构建数字资源共享平台框架可以较容易地构建数字资源共享平台的应用, 既满足了各个系统的共享与互操作, 又建立了一种地址解析系统, 将服务门户和数字资源提供者无缝连接起来。当然同时还有一些问题有待完善。比如: 建立元数据提供者的目录, 实现数据提供者基地址注册和更新维护的机制, 即由于数据提供者的数量较多, 并且基地址有变化的情况下, 手工维护就不能保证元数据的正常更新, 需要建立一种数据提供者基地址注册和更新维护的机制; 低成本的系统转换支持, 即作为元数据的拥有者, 如果为元数据的发布而单独建立数据发布系统, 将增加其系统运行的成本, 对数据维护和更新是不利的, 需要为内容拥有者提供内容管理与发布为一体的元数据发布模块, 使其很容易嵌入到内容发布系统中去, 以便降低系统转换的成本。

参考文献:

[1] ARTHA L, BROGAN A. Survey of digital library aggregation services[M] . Washington, D. C. : Digital Library Federation, 2003.

[2] 马文峰. 数字资源整合研究[J] . 中国图书馆学报, 2002, **28**(4) : 64-67.

[3] 肖珑. 元数据格式在数字图书馆中的应用[J] . 大学图书馆学报, 1999, **17**(4) : 23-25.

[4] 张晓林. 数字对象的唯一标识符技术[J] . 现代图书情报技术, 2001(3) : 8-11, 14.

[5] 林绮屏. 数字资源互操作协议 OAI 与 OpenURL 之比较研究[J] . 情报杂志, 2004, **23**(7) : 12-13, 16.

[6] 毛军. 我国数字图书馆标准规范建设: 数字资源唯一标识符的现状与发展[EB/OL] . (2004-05) . <http://cdls.nstl.gov.cn>.

[7] 牛振东, 朱先忠. 我国数字图书馆标准规范建设: OAI-PMH 协议应用指南[EB/OL] . (2004-05) . <http://cdls.nstl.gov.cn>.

[8] CNRI[EB/OL] . <http://www.cnri.reston.va.us/>.

[9] OAI-PMH[EB/OL] . <http://www.openarchives.org/>.

[10] OAICat[EB/OL] . <http://www.oclc.org/research/software/oai/cat.htm>.

(上接第 213 页)

[2] EISER M. Program slicing[J] . **IEEE Transactions on Software Engineering**, 1984, **210**(4) : 352-357.

[3] KOREL B, LASKI J. Dynamic program slicing [J] . **Information Processing Letters**, 1988, **29**(3) : 155-163.

[4] HALL R J. Automatic extraction of executable program subsets by simultaneous program slicing [J] . **Journal of Automated Software Engineering**, 1995, **2**(1) : 33-53.

[5] CANFORA G, CIMITILE A, LUCIA A De. Conditioned program slicing[J] . **Information and Software Technology**, 1998, **40**(11/12) : 595-607.

[6] GUPTA N, MATHUR A P, SOFFA M L. Automated test data generation using an iterative relaxation method: proc. of the 6th ACM SIGSOFT International Symposium on the Foundations of Software Engineering(FSE-6 SIGSOFT '98)[C] . Orlando: [s. n.], 1998:231-244.

[7] 赵瑞莲. 软件测试方法研究[D] . 北京: 中国科学院, 2001: 83-88.

[8] 王雪莲, 赵瑞莲, 李立健. 一种用于测试数据生成的动态程序切片算法[J] . 计算机应用, 2005, **25**(6) : 1445-1450.

[9] VIVEKANANDA M, VEDULA, JACOB A, *et al.* Program slicing for hierarchical test generation: proceedings of the 20th IEEE VLSI Test Symposium (VTS '02) [C] . [S. l.] : [s. n.], 2002.

[10] WONG W E, QI Yu. An execution slice and inter-block data dependency-based approach for fault localization: proceedings of the 11th Asia-Pacific Software Engineering Conference (APSEC '04) [C] . [S. l.] : [s. n.], 2004.

[11] 李必信. 程序切片技术及其在面向对象软件定量和软件测试中的应用[D] . 南京: 南京大学, 2000: 6-18.

[12] CHEN Zhenqiang, XU Baowen, YANG Hongji, *et al.* Test coverage analysis based on program slicing: IRI '03[C] . Las Vegas: [s. n.], 2003: 559-565.

[13] 黄光燕, 李晓维. 软件的变量完整性测试方法[J] . 计算机辅助设计与图形学学报, 2004, **16**(11) : 1584-1589.