学位论文全文数据库建设与探索

王 雁 凌 毅 李晨英 塔 娜

（中国农业大学图书馆 北京 100094）

【摘要】 结合本校博士、硕士学位论文全文数据库的建设实践，对学位论文远程提交、版权处理、元数据规范、数据采集与加工、信息发布等关键性环节进行了探索。

【关键词】 学位论文 全文数据库 研究生 【分类号】 G255

Exploration on Constructing Dissertation Full Text Database

Wang Yan Ling Yi Li Chenying Ta Na

(*Library of China Agricultural University, Beijing 100094, China*)

【**Abstract**】 This paper explores the key links of long－distance submission, copyright treatment, metadata standard, data collection and processing, information issuance in constructing the full text database about doctoral dissertations and master's theses of China Agricultural University.

【**Keywords**】 Dissertations Full text database Graduate students

1 学位论文的特点

博硕士学位论文与图书、期刊、报纸、会议记录、科技报告、专利、标准一样，是记载知识信息的一种重要的文献类型，与其它文献相比具有以下特点：

（1）独创性：博硕士学位论文、特别是博士学位论文，一般都涉足前人尚未研究过或尚未研究成熟的学科前沿性课题，因此博硕士学位论文是了解国内外科学研究与科技发展动态的重要信息媒介。

（2）学术价值：博硕士学位论文都是在某一学科有造诣的学者、专家的指导下完成，文献调查比较系统，研究方法与研究过程论述得比较具体，讨论分析具有独到的见解，具有很好的参考与借鉴价值。

（3）内部保存：多数学位论文是作为内部资料在学位授予单位图书馆或档案室以及国家规定呈缴的图书馆（如国家图书馆）中收藏保存，只能提供馆内浏览和复印，这种服务方式使学位论文的信息资源不能得到广泛的利用，其学术价值也不能得到充分的发挥。

2 建设学位论文全文数据库的必要性

随着计算机与信息技术的飞速发展，网络数字化文献信息资源的需求日益增大。为了提高学位论文学术资源的利用率、充分发挥学位论文的学术价值，高校及科研院所的图书馆、文献中心或资料室作为学位论文的长期收藏单位，有条件、也有责任将收藏的本单位博硕士学位论文进行数字化加工，为用户提供方便快捷的网上检索查询和全文服务。学位论文全文数据库的建立有助于研究生确定论文的选题和研究方向，避免与他人研究工作不必要的重复；帮助从事相近科研工作人员了解相关研究动态、借鉴有关的理论与方法，同时学位论文的网络公布使其研究和实验的结果受到更广泛的关注，这将进一步提高反盗版、反剽窃的能力，使作者的研究成果得到更好的保护；电子版学位论文的网上公布，还可以让更多的人了解和评判学位论文的水平，有助于促进指导教师精心指导学生、研究生努力写出高水平的学位论文。

3 国内学位论文全文数据库建设现状

近年来，国内一些大型文献信息机构凭借其作为国家定点学位论文收藏单位（如中国科技信息所、中国国家图书馆）、高等院校图书馆联盟机构（如全国高等教育文献保障系统CALIS）的优势，开展了一些大型学位论文文摘数据库的建设。国内一些高校、研究院所也纷纷利用本地资源自建了学位论文数据库（多数为题录或文摘型）。随着网络信息资源需求及利用率的提高，近两年来一些商业数据公司及重点高校正着重于学位论文全文数据库的建设，它们在数据来源、学科范围、全文电子文档格式、产品发布方式等方面都各具特色，下表是对三个有代表性的学位论文全文数据库基本情况进行的综合比较。

中国科技信息研究所是国家科技类学位论文的定点收藏单位，其收录学位论文的学科只限于理工科类，再加上在论文呈缴执行中的疏漏，因此万方数据公司的中国学位论文全文数据库有其不可避免的局限性。而清华同方公司则是通过与具有学位授予权的单位达成共建协议，加入其“中国优秀博硕士学位论文全文数据库”的共建单位目前约有300家，年增加学位论文约2万篇，但与全国目前有800多家博硕士学位授予单位、年产出约10万篇学位论文相比，其收录的论文数量是有限的。由此可见，开展大型学位论文全文数据库建设的关键

收稿日期：2003－02－27

问题是论文的来源，而其根本的问题则是著作权问题，上表中所列北大 2001 年学位论文的授权情况就从一个侧面反映了这个问题。

表 1

数据库名称	中国学位论文全文数据库	中国优秀博硕士学位论文全文数据库	北大博硕士学位论文全文数据库
制作单位	万方数据公司	清华同方光盘股份有限公司等单位	北大图书馆
数据量(2003－01)	10 万篇(年增加约 3 万篇)	3 万篇(年增加约 2 万篇)	约 869 篇
收录年限	2000 年至今	2000 年至今	2001 年
数据来源	中国科技信息研究所	全国 300 家博点授予单位	北大博硕士毕业生
学位范围	自然科学(理、工、农、医)	多学科范围	本校学科范围
电子文档格式	PDF	CAJ(使用专用浏览器)	PDF
发布方式	镜像	镜像、Web 版、光盘	校园网内部
授权情况	尚未完全取得版权	向提供数据的学位授予单位取得授权	向作者本人取得授权(2001 年的 1762 篇学位论文中有 869 位作者授权在校园网上即时发布全文)

4 我馆自建学位论文数据库概况

学位论文是各学位授予单位拥有自主版权的特色资源，有条件的图书馆应充分利用这一特色资源，开展学位论文全文数据库建设并在局域网内提供服务，为本系统的教学、科研提供学术参考和信息交流。我馆于1993 年就开始学位论文题录数据库的建设，到1998 年底共制作自1985 年以来本校所有的学位论文的题录型数据2571 条。1999 年起我馆作为成员馆每年向CALIS 提供CCFC 格式的本馆学位论文文摘数据，每年的数据量约为400 条。2001－2002 年，开展了本校硕博士学位论文电子版的收集和学位论文全文数据库的建设，在建设过程中对一些关键性环节和技术问题进行了探索与实践，本文将就我们工作中的一些思路和采取的措施加以总结，愿与同行共商。

5 学位论文的版权处理

5.1 学位论文的使用授权

关于学位论文的著作权归属及版权使用问题目前尚未有较明确的法定制度，1999 年教育部《高等学校知识产权保护管理规定》中第十三条规定："在高等学校学习、进修或者开展合作项目研究的学生、研究人员，在校期间参与导师承担的本校研究课题或者承担学校安排的任务所完成的发明创造及其它技术成果，除另有协议外，应当归高等学校享有或持有"，但这一条款明显不是针对著作权而言；另据第九条规定："为完成高等学校的工作任务所创作的作品是职务作品，除第十条规定情况外，著作权由完成者享有。高等学校在其业务范围内对职务作品享有优先使用权。作品完成二年内，未经高等学校同意，作者不得许可第三人以与高等学校相同的方式使用该作品"，此条款与《中华人民共和国著作权法》中的第十六条相吻合，但均未就学校或单位对作品优先使用权的业务范围加以明确限定，因此学位授予单位不经作者同意以法人身份将本单位学位论文发布权授予其它第三方的行为是否为法定允许值得探讨。

按照国内外惯例，由于学位论文中涉及研究和实验所取得的一切成果均是利用所在单位的物质技术条件或经费所取得的，所在单位对研究生学位论文具有收藏和优先使用的权利，图书馆可利用学位论文为促进学术交流而提供信息服务。因此硕博士毕业生在通过学位论文答辩之后必须呈缴学位论文印刷本，图书馆将其列为内部资料提供内部阅览及复印。我校规定研究生在通过学位论文答辩之后应向图书馆提交定稿学位论文的印刷本及全文电子文档，未提交者不予办理毕业离校手续或学位证书的领取。为了尊重和保护作者的著作权，我馆在收集电子版学位论文的同时，采取由作者及其导师签署"研究生学位论文版权使用授权书"的方式取得学位论文网络发布及传递的文字授权依据。授权书中要求研究生及其导师就学位论文的管理与使用、公开范围及方式、保密级别与公开时限等具体授权条款作出明确的认可或限定。

据我校 2002 年博硕士研究生及其导师对学位论文电子版使用授权的情况来看，我校有 20%学位论文电子版被授权在校园网上即时发布，未授权即时发布的学位论文主要有三种：一是保密级学位论文(涉及保密科研课题)，二是内部级学位论文(研究结果有待于申请专利或技术转让)，三是由于研究结果有待于在国内外专业期刊上投稿发表而推迟发布。授权中论文的保密时间一般为提交后1—5 年不等。图书馆将根据授权情况每年对限期已满的学位论文进行解密发布。

5.2 学位论文的著作权保护

随着我国科学研究向高、新、尖领域的发展，科研工作者的知识产权意识将越来越强，学位论文是对尚未公开的科研工作及成果进行描述和总结的原创作品，应给予相应的知识产权保护，特别是涉密学位论文的保护。从 2003 年起，我校将对学位论文的密级审定作严格的规定，图书馆在学位论文的复印、馆际互借以及数据库发布方面，也将采取严格的版权保护措施。对于电子版学位论文的版权保护措施主要是文档安全性保护及推迟发布的论文管理。文档安全性保护措施是将电子版学位论文格式转换为PDF 格式并对打印、文字编辑等功能加以限制；推迟发布的论文在授权时限内只在数据库中发布其题目、作者、联系方式及文摘等信息，在全文发布之前凡需要索取全文的用户可通过图书馆或自行与作者及导师取得联系。

6 数据库元数据规范

元数据标准是数据库结构的基础，是数据描述的准则。基

于当前国内外图书馆界对DC元数据标准的认可与使用，我馆以DC元数据为标准并参照高等教育文献保障体系CALIS《高校学位论文全文数据库建设参考（讨论稿）》，制定了本馆"学位论文全文数据库描述性元数据规范"。

从我馆规范的元数据的整体框架来看，基本上以DC元数据的15个基本元素项为基础，在制定具体元素项及描述内容时作了一些调整。具体如下：

（1）对于DC中与学位论文文献特性不是密切相关的元素项（如资源标识符Resource Identifier、来源Source、覆盖范围Coverage、出版Publication）暂时未采用；

（2）对于CALIS讨论稿中专设的"著者信息"元素项，我们认为其内容可分成两部分归入两个元素项中，其中"作者名称、学号、培养单位、院系、电话、E－mail、永久通讯方式"归入主要责任者Creator元素项中，"学科、专业、研究方向"则归入主题Subject元素项中；

（3）在资源形式Format元素项中，除了对论文的文档格式进行了描述，我们还对文档的大小及其显示页数进行了描述；

（4）关于日期Date元素项，与学位论文相关的日期有：论文答辩日期、学位授予日期和论文提交日期，究竟著录哪个日期、又以哪个日期为检索限定条件是值得探讨的问题。我们目前的做法为：三个日期均作著录，以学位授予日期为检索限定条件；

（5）为了便于今后数据交换以及世界范围的资源共享，在元数据中对学位论文的馆藏地址、学位授予单位、数据提供单位进行了描述，并提供了题名、文摘、关键词的英文录入。

7 数据库导航

建立与资源特性相适应的数据库导航体系，根据论文内容进行数据归类，从而把数据库资源组成为一个有条理的体系，将极大地提高用户检索查询的效率。学位论文的文献特性决定了其学科专业较强，因此按照学科专业对学位论文进行文献归类更具可行。我馆参照教育部颁布的最新学科专业、结合本校专业学科，制定出具有本校专业特色的学位论文全文数据库的分类导航体系，该体系将学位论文分为8个大类，下设28个一级学科，57个二级学科。需要注意的是随着学科专业的调整，该体系也需作相应的调整。TPI系统提供了数据库多重导航功能，学位论文数据库除了按学科设置导航体系，还可按中图法设置导航体系。多重导航体系的设置，为用户提供了灵活方便的导航选择。

8 数据采集

8.1 学位论文的远程提交

到目前为止，多数图书馆还是通过研究生提交软盘的方式收集电子版学位论文，只有北大、清华、武汉大学、西安交大、中山大学、上海交大及我校等几所高校的图书馆实现了网上学位论文远程提交。网上学位论文远程提交不仅可为研究生提交论文提供方便，又能提高图书馆工作人员的效率。我馆在系统开发人员的帮助下，实现了学位论文网上提交与数据采集相结合的系统功能。学位论文提交表单的内容主要是论文、责任者、专业的相关信息，表单中填写的信息按照数据库系统的程序指定到数据库元数据的相应内容中，从而省去了相应部分的元数据人工标引工作，提高了数据加工效率。研究生通过网络提交电子论文成功后，系统将回复显示"提交成功"页面。如果因文件过大或其它原因导致提交失败，可采取活动硬盘、光盘或图书馆匿名FTP方式提交论文。为了使研究生按照正确的方式和步骤完成电子版学位论文的提交事宜，我馆特制定并发布了电子版学位论文的提交程序，图示如下：

进入图书馆主页学位论文提交系统
↓
输入姓名学号
↓
按提示要求填写提交表单
↓
点击"浏览"选择论文文件所在路径并上传原文
↓
论文提交1－3天后在学位论文数据库中核查本人论文
↓
在图书馆网页下载并填写"学位论文版权使用授权书"（一式两份）
↓
到图书馆办理授权及验收手续

8.2 数据验收

我馆规定在研究生论文提交后3个工作日内对学位论文进行验收，合格后即在数据库中试发布，研究生通过校园网在数据库中核查本人论文无误后，来图书馆办理授权及验收回执。在学位论文提交验收的过程中，我们发现比较容易出现的问题有：

· 提交的电子版学位论文不完整，如有的缺封面，有的只有正文部分。

· 在低版本文字处理系统（如Word97）中编辑的论文在高版本系统（Word2000）中浏览时易出现版式错位现象。

· 若使用enter键强行分页，在版式改变时容易出现版式错位，建议使用"插入分页符"换页。

· 对于分章节撰写、编辑的学位论文，在将多个文件进行串接时出现页眉、页码错乱现象。有的研究生采取将多个word文件压缩打包的方式提交，但这样提交上来的仍旧是一些散文件。这种情况可以在不同部分及章节间插入分节符，在各自章节内分别设置不同的页眉和页码。

9 数据加工

（1）将学位论文电子文档转换为PDF格式 PDF文档格式是较为通用的网络文档格式之一，其优点有四：一是版式较为柔和美观；二是PDF文档具有翻页和书签导航功能，便于机上阅读；三是文件的大小较之同样的Word文档大幅度压缩，更加便于在线阅读；四是PDF文档具有一定的安全性，利用Acrobat相应的软件功能可进行文档安全性设置，例如：禁止打印、禁止复制、禁止更改等等，因此利用Adob公司的Acrobat软件将学位论文Doc文档转成PDF文档格式并进行安全性设置，更适用于学位论文的网络在线阅读。

（2）著录标引 由于学位论文网上提交表单中填写的信息可通过系统程序指定到数据库元数据的相应内容中，因此数据的著录标引工作重点是提交表单中没有的信息，如资源格式、权限管理、主题信息

以及全文链接等。

(3) 分类导航标引 通过建立分类栏目代码与元数据学科专业项的对应关系,TPI系统实现了自动分类标引功能,即按照研究生远程提交表单中所填写的专业,与数据库分类类目相对应,系统程序对数据进行自动分类标引。

(4) 数据备份 为了保障数据库的安全性,防止系统不稳定造成数据丢失,应及时进行数据备份。数据库数据可采取异地备份,源数据可采取刻录光盘保存。

10 数据发布

(1) 数据库检索功能 首先,在数据库元数据中选择、设定检索项,检索项应能体现学位论文文献特性。由于全文字段占用的数据空间特别大,而在一次检索中全文检索的结果往往与检索目标相关度较低,因此检索效率差,因此全文检索的意义不大,我馆学位论文全文数据库不提供全文检索功能。相比之下,选择题名(中英文)、作者、指导教师、专业、研究方向、文摘(中英文)、关键词(中英文)作为检索项较为适合学位论文检索需要。

其次,在检索技术方面,不同的数据库系统所支持的检索方式不尽相同。除了一般的简单检索与高级组配检索(布尔逻辑、截词符和位置算符控制)外,还应支持渐进检索(在检索结果中进行再次检索)、限制检索(如年代限制)以及全面检索(对所有检索项进行检索);检索匹配方式可有多种,如精确匹配、前方一致、中间包含和模糊匹配,提供多种检索匹配方式可建立灵活的检索策略,检索匹配方式是否完善取决于数据库系统功能。完善的数据库系统还可建立同义词、排除词词库,以提高检索的有效命中率。

总之,应尽可能提供灵活方便的检索功能,使用户在检索过程中可以根据需要不断调整检索策略,从而取得最佳的检索效果。

(2) 输出结果显示 学位论文全文数据库的输出结果显示方式可分为以下几个步骤:首先,为用户初步浏览提供检索命中结果的基本信息显示(如题名、作者、导师),命中结果应能按论文发表时间或相关度(以检索词出现频率进行量化)进行排序,并提供检索结果的保存、打印以及E-mail发送;然后根据用户对目标论文的选择点击,显示目标论文的有关详细信息(包括文摘、馆藏信息等);最后对于授权在校园网公开的学位论文提供全文电子文档在线浏览与下载(下载的PDF文档不提供编辑权限)。对于保密或内部管理的学位论文,用户可在数据详细信息中查询到该学位论文的网上公开时间及纸本论文馆藏索取号。

(3) 信息发布范围 学位论文全文数据库的网络信息发布应遵守学位论文作者在版权使用授权书中所授权的公开范围及年限,对于作者同意公开全文的学位论文或经数字化转换、回溯加工的馆藏印刷本学位论文,一般只宜在校园内或图书馆内予以全文发布。

参考文献:

[1] http://www.lib.pku.edu.cn/whatsnew/news/xwlw_2001.htm
[2] http://mirror.wanfangdata.com.cn:98/
[3] http://www.cnki.net/

(上接第62页)

Logic:Null空值或Or、And逻辑连接符Comparetype:布尔值,决定是与一个常量比较,还是与另一个字段比较

Operation:大于、小于、等于、不等于等关系操作符

Value:一个常量值,或某个字段名

3.4 处理事件逻辑

Inputtree和Designtree控件作为具体的用户界面对象,具备丰富的事件机制。如删除节点、新增节点、选择和输入一个值、点击、焦点转移等事件,既涉及到界面本身的改变,也涉及到相应数据的更新。为了使之同步而不致出现逻辑上的错误,一般需要在每个必要的事件过程中同时处理界面的变化和数据的变化,以实现"对象一关系映射"的一致性。在处理与数据有关的持续化操作时,主要是通过调用构建的Treebone对象和Inputconstraint对象的接口属性与方法来加以处理:根据控件上的事件变化修改数据表,以及从数据表读取数据以决定下一步控件界面应如何改变。

4 小 结

对象持续化是软件编程中常见的需求。"对象/关系映射"是处理对象持续化的一种重要方法。由于MIS应用开发中几乎不可避免都要用到关系数据库系统,因此采用"对象/关系映射"实现对象持续化是一种自然的选择。尽管关系表不是天生为描述对象性质而设计的;但只要设计合理,它可以具有足够的描述能力,满足大多数应用的需要。

参考文献:

[1] 刘圣才,李春葆编著. Visual Basic 6程序设计导学. 北京:清华大学出版社,2002,1
[2] Ed Roman 著,王进亮等译. 精通EJB. 北京:电子工业出版社,2002,1
[3] 李智慧,秦 成编著. C++Builder 4.0从入门到精通. 北京:清华大学出版社,1999,8
[4] Billy Hollis, Rockford Lhotka 等著,康 博译. VB.NET程序设计教程. 北京:清华大学出版社,2001,10