

A Study
of
Library
Bibliographic
Search

图书馆

文献搜索研究

杨新涯
著



重庆大学出版社
<http://www.cqup.com.cn>

A Study
of
Library
Bibliographic
Search

图书馆

文献搜索研究

杨
新
涯
著

重庆大学出版社

图书在版编目(CIP)数据

图书馆文献搜索研究/杨新涯著. —重庆:重庆
大学出版社,2015.5

ISBN 978-7-5624-8978-8

I. ①图… II. ①杨… III. ①情报检索—研究 IV.
①G252.7

中国版本图书馆 CIP 数据核字(2015)第 068703 号

图书馆文献搜索研究

杨新涯 著

策划编辑:林佳木

责任编辑:陈 力 版式设计:林佳木

责任校对:谢 芳 责任印制:赵 晟

*

重庆大学出版社出版发行

出版人:邓晓益

社址:重庆市沙坪坝区大学城西路 21 号

邮编:401331

电话:(023) 88617190 88617185(中小学)

传真:(023) 88617186 88617166

网址:<http://www.cqup.com.cn>

邮箱:fxk@cqup.com.cn(营销中心)

全国新华书店经销

重庆川外印务有限公司印刷

*

开本:787×1092 1/16 印张:18.25 字数:263千

2015 年 5 月第 1 版 2015 年 5 月第 1 次印刷

ISBN 978-7-5624-8978-8 定价:48.00 元

本书如有印刷、装订等质量问题,本社负责调换

版权所有,请勿擅自翻印和用本书

制作各类出版物及配套用书,违者必究

今年春节期间，国民远赴日本国抢购马桶盖和大米，成了新闻热点。在沸沸扬扬的网络争论中，中科院的专家也出来表态，说今后将在“高精尖”研究的同时，加大力量研究公众所需要的“小玩意”。

在图书馆学领域，又何尝不是这样，热衷于研究一些吸引眼球的方向，不断诞生很多新词、热词，不断将其他领域的研究“交叉”到图书馆之中，但是一些真正有助于文献管理和服务的基础研究，已经被忽视很久了。坦白地说，也许是我愚笨的缘故，现在一些图书馆学、情报学的研究论文，根本就看不懂，也不清楚这些研究对于图书馆有什么作用。但我宁愿相信，这是上面抢购国外马桶盖的那个原因——我们有能力让中国人去太空漫步，却连日常生活的必需品都做不到。

目录学，就是被目前图书馆学研究所忽视的一个最基础的、最根本的方向，却又是对图书馆管理和服务工作，对读者利用图书馆最为重要的研究方向。而恰恰在近十年来，目录学需要研究和突破的必要性超过以往任何时候，因为我们

处在一个极其复杂的文献资源环境中。

以重庆大学图书馆为例,从2000年开始购买用于网络服务的数字文献资源,数量越来越多,种类越来越复杂。在当时的五年计划中,我“大胆”地将未来5年的存储容量规划到2TB,结果5年后是20TB。而现在,在大量的远程访问数据库的背景下本地存储也超过200TB。除纸质文献之外,图书馆收藏了光盘、磁性材料的纯文本文献,也有图像、声音、影像等更多直观的媒体文献。常用的数字文献出版物类型都有入藏:数字图书、数字期刊、全文数据库、网络报纸、电子地图、数值型数据库、文献管理软件系统等。互联网和云计算的推广又加速了图书馆文献组织的复杂性,越来越多的文献资源根本就不在本地服务器上,而是按年或者永久购买远程使用权。而且,互联网的海量免费资源,也开始进入图书馆的服务范围。数字文献种类和数量的急剧增长,给图书馆行业带来了相应的问题:元数据标准的不统一、数字格式的多样性、独立的管理和服务系统等;众多的数据库以“信息孤岛”的方式而存在,相互之间并没有数据关联,检索也是各自独立的,因此图书馆不得不提出“统一检索平台”或者“知识发现系统”的解决方案,期望通过构建一个从各个数据库抽取出来的元数据库,解决不同数据库之间的集中检索问题。但是从这些年的实践来看,现有的方案都有各种各样的问题,而依附于统一检索系统的各类文献服务,更是没有有效开展起来。

这些,都是图书馆目录学需要研究的“民生”问题。

因此我在导师程焕文教授的指导下,用了大概两年的时间完成了文献搜索的研究,提出以整合图书馆服务为目标,重建纸质文献和数字文献的标准控制体系,形成属于图书馆的二次文献数据仓储,在此基础上构建整合图书馆服务的搜索平台,使之成为图书馆的第一门户,有效提升图书馆的文献服务能力,并积极通过数据服务拓展图书馆的文献服务范围,提高图书馆社会影响力。

个人认为这项研究对于图书馆目录学具有实际的利用价值:

第一,在数字图书馆日益发展的今天,图书馆面临纸质文献和数字文献各自独立服务的现状,通过文献搜索的基本理论的梳理,应以

文献搜索的构建来完善现代图书馆目录学,包括概念、原则和方法,纸质文献和数字文献的标准控制体系、搜索平台等,以此作为图书馆服务的基础,重建复合型图书馆的文献管理、服务模型,改变当前仅以纸质图书的 OPAC 为核心的传统模式,提高复合型图书馆的文献资源管理和服务能力。

第二,研究提出了文献搜索的三大基本特征:元数据搜索,整合文献,整合服务。按照这三大基本特征,明确图书馆建设文献搜索的系统边界、系统需求设计的九个需求设计要点,在元数据仓储的基础上拟定元数据存储、描述与控制、输入与输出的三层标准控制体系,倡导构建“整合搜索+垂直搜索”相结合的、以用户为核心的图书馆文献搜索平台,在系统架构设计中,坚持了以用户为核心、系统整合和元数据仓储三个原则,并利用 Web2.0 技术,增加读者的参与和互动,改善读者的搜索使用体验,为读者提供“一站式”服务。

第三,研究了文献搜索拓展到其他应用系统中的可行性,提出应尽可能多地整合图书馆服务进入搜索中,通过系统接口的方式支撑各行各业在业务工作中对于知识的需求,主动将文献搜索的知识库推送到文献共享系统、手机图书馆、在线科研系统中,为其他业务系统提供实时的知识支撑和服务支撑,变被动服务为主动服务。

第四,图书馆应探索全新的、整合传统文献资源和数字文献资源的 RMS (Resources Management System) 系统,将文献搜索构建为图书馆第一门户,尝试突破现有的资源管理系统的设计思路,构建对立统一的图书馆管理模型,将图书馆管理分为“资源”和“服务”两方面,实施图书馆流程再造的计划,提高现代图书馆的管理和服务水平,提高图书馆的社会价值。

尽管一些人习惯使用百度和谷歌,但我从来没有质疑过图书馆构建文献搜索系统的必要性,当这些搜索引擎以几何级数增长的时候,查全率和查准率已经成为难题,我们需要的内容正被越来越多的垃圾信息所隐藏,因此这些搜索引擎也在日益重视垂直搜索的构建。而应用于学术研究和公众阅读的文献搜索,理当由图书馆来研究和推行。

这是一项任重道远的事业，在完成研究的两年后，业界倡导的“知识发现”在图书馆流行起来，却依然被商业模式所主导。单凭一己之力，实践起来是那么艰难，但我始终坚信，哪怕是一点点的进步也是进步，就这样继续前行吧。

2015 年 3 月 10 日
重庆大学·可也斋
新涯自序

目 录

第一章 绪论	1
第一节 有意义的前期研究与实践	3
一、读者时代的需求	3
二、图书馆 2.0 研究和实践	4
三、与读秀学术搜索的服务整合	11
第二节 文献搜索的实践现状	19
一、国内外整合文献资源的主要方案	19
二、元数据收割模式的整合搜索现状与不足	22
第二章 文献搜索理论与方法	35
第一节 概念的分析	35
一、检索和搜索	35
二、文献搜索的三个基本特征	38
三、文献搜索属于目录学范畴,是图书馆工作的核心内容	39
四、传统目录学对数字资源组织方法不足	40
五、现代目录学需要完善的方法论	46
第二节 从 OPAC 到文献搜索	49
一、文献搜索在现代化图书馆系统中的位置	50
二、文献搜索是图书馆第一门户	53
三、系统边界问题	54
四、未来图书馆目录组织的基础架构: RMS	61
五、文献搜索的价值	64
第三节 整合搜索与垂直搜索	68
一、垂直文献搜索概述	69
二、以读者为导向的垂直文献搜索需求分析	70

第三章 文献搜索的元数据研究	80
第一节 元数据概述与面临的困难	80
一、数字文献大多采用企业标准,其文献元数据具有封闭性	81
二、数字文献部分提供 MARC 或标准接口	83
第二节 元数据描述与存储的主要标准	83
一、机读目录	84
二、都柏林核心 (Dublin Core) 元数据集	86
三、DOI 和 CALIS-OID	90
四、OpenURL	94
第三节 元数据质量保障与输入和输出	95
一、元数据质量保障	95
二、元数据的输入与输出	99
第四节 元数据标准控制体系与管理	108
一、元数据标准控制的“三原则”	109
二、元数据标准控制的框架	110
三、《元数据应用纲要》编制与实例	112
四、元数据管理器	115
第四章 图书馆服务的整合	122
第一节 对 OPAC 整合服务的调研	122
一、斯坦福大学图书馆	123
二、纽约公共图书馆	132
三、南京大学图书馆	139
第二节 整合的图书馆文献服务	143
一、图书馆服务的分类和分级	144
二、可供整合的文献服务需求分析	147
三、服务整合的方式与方法	151
第五章 文献搜索平台的构建	157
第一节 典型搜索引擎的特点和模型分析	157
一、电子商务类搜索引擎	157
二、图书馆行业领先的搜索引擎	165

三、主要搜索引擎的经验	176
第二节 文献搜索平台的系统架构	178
一、系统架构三原则	178
二、需求设计要点	179
三、系统整体架构	184
第三节 文献搜索平台的实现	187
一、软件架构的选择	188
二、技术框架	192
三、全文搜索引擎	194
四、其他重要信息技术与安全	198
第六章 实证研究：弘深文献搜索设计概要	207
一、系统概况	207
二、基本结构图	210
三、系统需求概要	210
第七章 结语：文献搜索的拓展与未来	245
第一节 文献搜索在共享系统中的应用	245
一、文献共享是共知的基础	245
二、整合资源为重点的共享体系：以 CALIS 的 e 读为例	247
三、整合服务为重点的共享体系：重庆科技文献资源共享平台 ...	252
第二节 “让图书馆无处不在”	255
一、以文献搜索支撑移动图书馆	255
二、文献搜索的数据服务	262
参考文献	267
后记	274

图表目录

图 示

- 图 1-1 基于 J2EE 的图书馆 2.0 的五层架构体系
- 图 1-2 图书馆知识搜索 LKS 的元数据流程
- 图 1-3 图书馆知识搜索 LKS 的服务整合实例
- 图 1-4 万方学术搜索系统的架构
- 图 2-1 传统图书馆的文献服务模型
- 图 2-2 复合型图书馆时期的服务模型
- 图 2-3 文献搜索在现代图书馆系统中的位置
- 图 2-4 图书馆文献搜索的系统环境图
- 图 2-5 和谐统一的图书馆管理模型
- 图 2-6 图书馆垂直文献搜索的基本类别
- 图 2-7 学位论文的基本分类导航
- 图 2-8 按二级学科、专业条件搜索的搜索结果页面
- 图 3-1 元数据管理器的系统架构
- 图 3-2 图书馆元数据收割模式
- 图 4-1 SULAIR 的服务网站
- 图 4-2 SearchWorks 的搜索界面
- 图 4-3 SearchWorks 搜索流程 1: 简单搜索窗
- 图 4-4 SearchWorks 搜索流程 2: 搜索结果
- 图 4-5 SearchWorks 搜索流程 3: 搜索结果的具体款目
- 图 4-6 SearchWorks 搜索流程 4: 具体文献的详细信息
- 图 4-7 SearchWorks 搜索流程 5: 精练搜索的导航树
- 图 4-8 SearchWorks 搜索流程 6: 搜索结果输出
- 图 4-9 斯坦福大学图书馆的 OPAC 系统 Socrates
- 图 4-10 斯坦福大学图书馆的 Databases and articles 搜索
- 图 4-11 斯坦福大学图书馆的 XSearch 搜索系统
- 图 4-12 斯坦福大学图书馆课程文献搜索

- 图 4-13 NYPL 的搜索体系
- 图 4-14 NYPL 的一键搜索窗口
- 图 4-15 NSPL 的一键搜索功能
- 图 4-16 搜索结果的精练
- 图 4-17 NYPL 的 Web2.0 应用
- 图 4-18 图书资源的展示页面
- 图 4-19 图片资源的展示页面
- 图 4-20 NYPL 的电子图书搜索
- 图 4-21 NYPL 的数据库搜索
- 图 4-22 汇文 OPAC 的详细页面
- 图 4-23 汇文 OPAC 系统的可视化分析
- 图 4-24 图书馆服务方式的大致分类
- 图 4-25 图书馆门户的架构
- 图 4-26 线性图书馆管理模型
- 图 5-1 当当网
- 图 5-2 当当网搜索结果页面
- 图 5-3 Web of Science 的表单检索
- 图 5-4 Web of Science 数据库的检索结果展示
- 图 5-5 检索结果的导出
- 图 5-6 “分析检索结果”功能页面
- 图 5-7 “思源探索”系统功能图
- 图 5-8 “思源探索”二级页面截图
- 图 5-9 “思源探索”三级页面截图
- 图 5-10 文献搜索的三层架构
- 图 5-11 文献搜索的业务流程
- 图 5-12 文献搜索的 J2EE 标准技术框架
- 图 5-13 采用 Ajax 技术的系统界面
- 图 5-14 统一认证平台功能结构
- 图 5-15 系统安全保障体系
- 图 5-16 CA 平台功能组成图
- 图 6-1 弘深搜索的系统结构图
- 图 6-2 弘深搜索的功能结构图

- 图 6-3 “我的专题库”参考界面
- 图 6-4 弘深搜索首页参考图
- 图 6-5 期刊论文搜索结果样式
- 图 6-6 新增资源的功能流程图
- 图 6-7 数据导入功能流程图
- 图 6-8 作者数据库的流程图
- 图 7-1 e 读的体系结构
- 图 7-2 重庆大学图书馆的 e 读集成模式
- 图 7-3 e 读的访问和使用流程
- 图 7-4 学科网角色与功能的用例图

表 格

- 表 1-1 图书馆 1.0 和图书馆 2.0 的比较
- 表 1-2 图书馆知识搜索 LKS 的系统整体框架
- 表 1-3 图书馆知识搜索 LKS 的元数据收割对象
- 表 1-4 图书馆知识搜索 LKS 的中文元数据统计
- 表 1-5 几种主要元数据收割模式的搜索系统比较
- 表 2-1 主要的文献元数据基本处理单元边界
- 表 2-2 Exif 元数据信息实例
- 表 3-1 文献搜索的元数据标准控制框架
- 表 5-1 Web of Science 功能图
- 表 5-2 文献搜索的主要业务子系统
- 表 6-1 弘深搜索的功能列表
- 表 6-2 元数据管理工作量统计界面参考图
- 表 6-3 垂直搜索管理工作量统计界面参考图
- 表 6-4 搜索词界面参考图
- 表 6-5 年资源增长情况统计界面参考图
- 表 6-6 月资源增长情况统计界面参考图
- 表 6-7 资源分布统计界面参考图

绪 论

数字图书馆是图书馆发展的新形式,其大规模应用和发展也不过是近十年的时间。在这十年时间内,数字图书馆建设和研究可以用资源数字化、各种信息技术、网络化服务、知识产权、标准规范、业务模式等标签进行总结^①,形成了传统图书馆和数字图书馆共同发展的复合型阶段。在这个阶段,如何实现传统纸质文献和数字文献的统一揭示,并形成全新的图书馆服务模式成为当务之急。

随着数字文献资源和网络服务的增加,文献数据越来越多,数据结构越来越复杂,不同的数字文献资源有其各自的物理和逻辑结构,具有各自不同的检索系统和服务平台,检索方式、检索手段各式各样,使得信息量剧增但进一步导致了文献资源的无序和分散,用户的检索负担、用户的教育负担等日益加重。可以说,没有哪一个图书馆能够明确此时此刻,图书馆拥有的纸质文献资源与数字资源的具体数量,读者也需要在馆藏目录和多个数据库中检索才能得到收藏信息,而在以纸质文献为主的实体收藏时期,通过文献目录或 OPAC 系统就可以获知,从这个意义上来说,目前图书馆的文献资源管理能力与原来相比有所下降,这也会影响到图书馆的服务水平。

问题的根源在于,图书馆作为文献收集、整理、服务的社会机构,进入数字图书馆时代后,对于文献整理这个中间环节的理论和方法不足:如对数字文献资源的揭示深度,元数据标准定义和选取、数字资源组织、数字资源标引等基本业务,以及通过技术手段,纳入图

① 刘炜.再续数字图书馆十年的辉煌[J].中国图书馆学报,2008(3):84-85.

图书馆日常业务流程的数字文献资源工作规范。图书馆已经具有完整的围绕 OPAC 的传统文献服务模型:通过采访、编目、典藏,形成图书馆中央数据库,在此基础上构建了 OPAC 系统,读者通过对 OPAC 系统的查询,进而开展对图书馆的利用。但是数字图书馆时代,大多数图书馆对于各种不同标准、不同平台、不同类型的数据库,仅仅采用列表方式,提供在图书馆的主页上,形成“OPAC + 数据库列表”的文献服务方式,读者需要一个个数据库去检索和获取自己需要的文献,随着数据库的增多,图书馆的培训和教育负担、读者的检索负担也就越来越沉重,这与数字化的便利性相去甚远,这就是为什么数字图书馆在新时代,文献保障的能力大幅度提高,而文献服务水平并没有显著提高的根源,很多基于网络的学科服务、科研支撑、在线学习、社会服务等工作无法进一步开展。

2005 年 8 月美国威斯康星大学图书馆就提出“One System, One Library”的发展战略^①,但时至今日,很多图书馆的应用系统不仅没有越来越少,反而越来越多,信息孤岛现象日益严重。研究试图通过文献搜索的构建,将围绕在图书馆周围的纸质资源、数字资源、各种系统、各种标准系统地联系起来,有效整合图书馆文献资源和服务,形成图书馆系统的整体,由此提高现代图书馆的管理和服务水平。

如何通过理论和方法的研究,以及信息技术手段,实现互联网时代图书馆数字资源和纸质资源的整合,图书馆文献服务的整合,拓展图书馆对外应用,以有效提升图书馆管理和服务水平,是研究的主要目的。研究在图书馆 2.0 实践的基础上^②,以用户为核心,提出升级 OPAC (Online Public Access Catalogue, 在线公共目录查询系统)为文献搜索,整合图书馆的文献资源,整合图书馆的文献服务。

① 叶新明. 美国 ERMS 产品的分析与启示[J]. 中国图书馆学报, 2007(4): 58.

② 杨新涯, 彭晓东. 2.0 的图书馆[M]. 广州: 中山大学出版社, 2011.

第一节 有意义的前期研究与实践

一、读者时代的需求

近现代图书馆一直倡导以读者服务为核心构建图书馆系统,摒弃原来的“重藏轻用”思想,但是收效甚微。新世纪以来,随着个性化时代的来临,这种思想越来越受到重视。2003年,程焕文教授在“面向21世纪课程教材”《信息资源共享》中,根据各国图书馆信息资源共享的历史与现状、理论与实践,总结归纳出四个基本定理:定理一,一切信息资源都是有用的;定理二,一切信息资源都是为了用的;定理三,人人享有自由平等地利用信息资源的权利;定理四,用户永远都是正确的。这四个基本定理揭示了人类社会信息资源共享的基本观念、普遍原则和社会价值,是信息资源共享的理论核心和实践准则^①。

定理四“用户永远都是正确的”揭示信息资源共享服务的用户观念和基本原则,引起了图书馆界的广泛关注和争论,也让这个定理进入很多图书馆员思想深处。可以说,如何去实践和运用这个信念,决定了图书馆管理和图书馆服务的发展方向、路线和结果,也说明新世纪图书馆行业对于读者权利的重视^②。越来越多的图书馆开始尝试为读者提供更多的个性化服务:定制收藏、个人门户、学科专题文献推送、手机图书馆定制等。这些个性化服务也为随之而来的图书馆2.0的起源和发展奠定了基础。

但是在图书馆实际工作中,这些个性化服务无法解决读者利用图

① 程焕文,潘燕桃.信息资源共享[M].北京:高等教育出版社,2004:18.

② 程焕文.关于改变图书馆学研究立场的思考——从“用户永远都是正确的”说起[J].中国图书馆学报,2008(3):89-93.

书馆的根本诉求:文献需求。图书馆文献资源量得到大幅度增加,实现了365天全天候文献服务,且利用互联网将服务范围拓展到全球,在时间和空间方面的文献服务能力的确是革命性提高。但是在文献搜索的便利性方面,反而有所降低,读者不得不在越来越多的数据库中,搜寻自己所需要的文献,很多时候由于信息不对称,读者不清楚自己所需要的文献在哪一个数据库中而造成漏检。查全率与查准率成为了概念,而非可供检验的数据,因为图书馆不知道分母是多少。

提升文献服务的便利性以解决上述问题,就要求图书馆为读者构建类似谷歌、百度的搜索引擎,能够搜索图书馆收藏的全部文献,也就是本研究提出的整合资源的文献搜索。要实现这个目标,需要三个条件:一是统一的图书馆建设和发展规划,从顶层设计入手,尽量避免信息孤岛的产生,保证系统的一致性和兼容性;二是数字图书馆标准体系,数据是系统的核心,标准化是建设和发展的前提,也是文献资源共享的基础,尽管目前已经制订了数字图书馆相关的多种标准,标准规范的系统性建设已经得到很大的缓解,取得了一些实实在在的研究进展,然而却陷入刘炜在《再续数字图书馆十年辉煌》中所说的“似乎标准规范越多,离数字图书馆建设的实际越是遥远”的悖论^①,如何使各种各样的标准从图书馆馆员和用户的视角中消失,使其隐含在图书馆各种应用和业务流程之中,提高规范管理能力、水平,才是制订和颁布各种标准的最终目的;三是可持续的技术和管理保障措施,任何系统的稳定发展,都离不开持续创新的新技术的保障,以及相应的管理措施。

二、图书馆2.0研究和实践

笔者自2001年开始,从事数字图书馆的建设和管理工作,亲身经历了我国高校图书馆快速发展变化的十年。从2006年开始进行图

^① 刘炜.再续数字图书馆十年的辉煌[J].中国图书馆学报,2008(3):84-85.