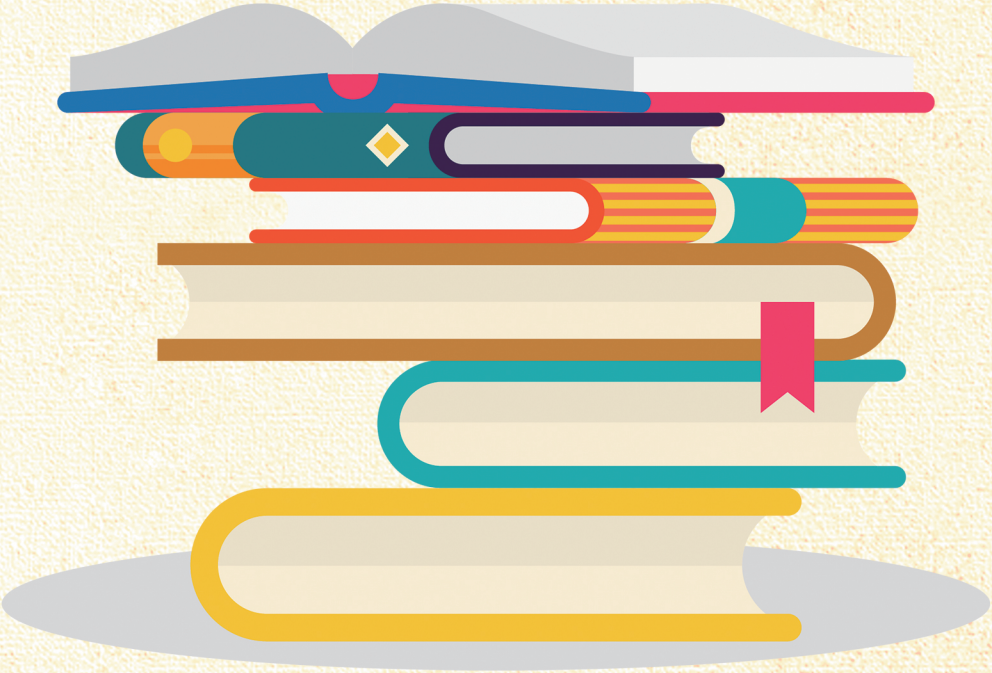
大数据环境下

图书馆文献信息资源建设与利用

刘付霞　著



吉林人民出版社

大数据环境下图书馆文献信息资源建设与利用

刘付霞◎著

吉林人民出版社

图书在版编目（CIP）数据

大数据环境下图书馆文献信息资源建设与利用 / 刘付霞著. -- 长春：吉林人民出版社，2019.11

ISBN 978-7-206-16547-4

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①数据处理－应用－院校图书馆－图书馆工作－研究 Ⅳ. ①G258.6-39

中国版本图书馆CIP数据核字(2019)第269834号

DaShuJu HuanJing Xia TuShuGuan WenXian XinXi ZiYuan JianShe Yu LiYong

大数据环境下图书馆文献信息资源建设与利用

著　　者：刘付霞
责任编辑：郭　威
封面设计：飒　飒
吉林人民出版社出版发行（长春市人民大街7548号 邮政编码：130022）
印　　刷：长春市昌信电脑图文制作有限公司
开　　本：880mm×1230mm　　1/32
印　　张：3.5　　字　　数：54千字
标准书号：ISBN 978-7-206-16547-4
版　　次：2019年11月第1版　　印　　次：2019年11月第1次印刷
定　　价：35.00元

如发现印装质量问题，影响阅读，请与印刷厂联系调换。

前　言

随着互联网、传感网络、社交网络、智能移动设备等技术的发展，数据呈现爆炸式增长，数据结构类型也复杂多样，海量数据充斥着每个角落，种种特征表明大数据时代已经到来。大数据时代，用户信息需求越来越个性化、专业化，面对丰富的信息资源，用户的信息选择障碍使其难以找到精准度高的需求信息，这加速了个性化信息服务的发展。个性化信息服务的理念是满足用户个性化信息需求，主动为用户提供个性化信息服务。

作为教师和学生获取信息的重要渠道，怎样在信息海洋中利用大数据挖掘技术找到满足用户个性化需求的信息，为用户提供真正意义上的个性化信息服务，是高校图书馆当下面临的重要挑战。将先进大数据技术引入高校图书馆个性化信息服务中，借助已有的个性化信息服务系统平台，结合大数据技术处理流程，构建基于大数据的高校图书馆个性化信息服务系统模型，为图书馆文献资源建设提供了新的帮助。

本书以“大数据环境下图书馆文献信息资源建设与利用”为选题，从文献信息资源的概念与特点、网络文献信息资源以及大数据资源平台的整合与图书馆信息管理切入，探讨图书馆文献

信息资源建设、图书馆文献信息资源的检索与利用以及大数据环境下图书馆文献信息资源的管理策略与创新。本书一方面内容细致、严谨、完善，阐述了大数据、个性化信息服务以及高校图书馆个性化信息服务相关理论，在厘清大数据与高校图书馆个性化信息服务概念和特征基础上，找到两者的契合点，着眼于高校图书馆个性化信息服务发展趋势，发现两者存在结合的可能性，这对大数据环境下图书馆文献信息资源建设与利用具有非常重要的意义。另一方面，本书文字通俗易懂，视野开阔。在撰写过程中，笔者分析了高校图书馆个性化信息服务，并结合大数据技术提出改善策略，尽量将理论性较强的内容变得通俗易懂，力图层次清晰，语言流畅，文字精练。

笔者在撰写本书的过程中得到了许多专家学者的帮助和指导，在此表示诚挚的谢意。由于笔者水平有限，加之时间仓促，书中所涉及的内容难免有疏漏之处，希望各位同行、专家、老师多提宝贵意见，以便笔者进一步修改，使之更加完善。

刘付霞

2019 年 4 月

目录 Contents

第一章　绪　论

第一节　文献信息资源的概念与特点 2
第二节　网络文献信息资源概述 4
第三节　大数据资源平台的整合与图书馆信息管理 14

第二章　图书馆文献信息资源建设分析

第一节　文献信息资源建设的内容与任务 20
第二节　文献信息资源建设的原则探究 41

第三章　图书馆文献信息资源的检索与利用

第一节　图书馆文献信息资源检索概述 52
第二节　图书馆文献信息资源检索技术探究 59
第三节　图书馆文献信息资源管理与利用研究 85

第四章　大数据环境下图书馆文献信息资源的管理策略与创新

第一节　大数据环境下图书馆文献信息资源建设的策略 90
第二节　大数据环境下的图书馆文献资源建设模式的创新 95

参考文献 101

第一章
绪　论

文献信息资源是以文字、图形、符号、声频、视频等方式记录在各种载体上的知识和信息资源。它包括图书、连续出版物（期刊、报纸等）、小册子以及学位论文、专利、标准、会议录、政府出版物等。本章重点探讨文献信息资源的概念与特点、网络文献信息资源以及大数据资源平台的整合与图书馆信息管理。

第一节　文献信息资源的概念与特点

一、文献信息资源的概念解读

资源，一般指天然资源。文献信息资源是相对于天然资源的一种社会智力资源，是物化了的知识财富，它是人们迄今为止收集、积累、贮存下来的文献信息的总和。文献信息资源作为一种宝贵的智力资源同水资源、矿产资源等自然资源一样是人类文明发展必不可少的条件。一个国家文献信息资源的多少及其存取水平是衡量该国文明水准和经济、文化、科学技术等综合国力的重要标准。文献信息资源的开发利用程序直接影响到社会的发展与进步。由于历史、经济、文化等诸方面的影响，不同国度的文献信息资源多寡不均，同一个国家不同地区的文献信息资源亦多寡不一。一般情况下，发达国家和发达地区的文献信息资源比较丰富[①]。

① 刘秉文，王志国，李志勇 . 现代文献信息资源建设 [M]. 呼和浩特：内蒙古人民出版社，2008.

二、文献信息资源的特点分析

（1）再生性。文献信息资源不像自然资源（如煤、森林等）那样会由于开发和利用的深入而逐渐枯竭，它具有再生性，可以多次反复地使用，还会随着开发利用程度的提高，反过来促进知识增值和文献信息资源建设工作的发展。可以说文献信息资源是取之不尽、用之不竭的再生性宝贵资源。

（2）可建性。文献信息资源是人类创造的一种信息智力资源，人们可以通过规划、采集、选择、布局等手段对其进行建设、改造和优化，使其成为方便人们利用的有序整体。

（3）积累性。文献信息资源的多寡不是先天固有的，而是后天不断积累的结果，今天丰富的文献信息资源离不开历史上各个时期保存下来的各类文献信息资源，它是古代私人藏书家、官方藏书楼及近现代图书馆、各类文献收藏机构保存下来的人类文明的集合。

（4）冗余性。文献信息资源并非各单位文献简单地相加，庞杂、雷同的文献堆积既不会增加文献信息内容的含量，也不会成为体系完备、功能良好的文献信息资源系统。文献信息资源建设的具体任务之一就是要把那些重复交叉以至过时无用的文献——冗余文献进行剔除，否则，有可能造成文献信息通道的阻塞，给用户带来困难。

（5）共享性。文献信息的共享性是由文献信息的社会占有性决定的。文献信息一旦产生并公布于世，社会公众就有了平等

利用的机会。文献信息以各种方式出版发行的目的，从根本上说，是为了让更多的人去利用它。文献信息的这种共享性给人们开展文献信息资源的共建、共享工作提供了理论依据。

（6）价值潜在性。文献信息资源的价值实质是文献信息载体所含知识内容的价值。在被开发利用之前，这种价值潜在于载体之中，不为人们所见；在被开发利用之后，这种价值间接体现于某种产品、某种成果、某种思想、观念或行为之中，具有隐现性。知识含量越多的产品其价值越高，文献信息资源被开发利用得越好，物质成果和精神成果就越丰富。文献信息资源的价值是随着文献信息资源的开发程度而发生变化的。随着知识经济时代的到来，文献信息资源的价值必将被越来越多的人所认识。现在许多发达国家已把文献信息资源同能源、材料并列看作社会经济和科学发展的“三大支柱”之一。随着以知识为依托的知识经济时代的到来，文献资源的作用将更为突出。

第二节　网络文献信息资源概述

一、网络信息资源概述

在数字化的新时代，因特网以其自身优势为当今社会提供了一个全新的世界。网络上全方位的交流，使网络信息资源应运而生，且丰富多样。网络信息资源为图书馆信息资源注入了新的

血液，为图书馆在网络时代发展提供了新的契机。网络信息资源形式多样，有网络数据库、网络出版物（图书馆网上公共目录）OPAC、电子论坛、动态信息等。[①]

（一）网络信息资源的定义

网络信息资源虽然已被越来越多的人所接受，但目前对网络信息资源并没有统一的定义。网络信息资源就是通过计算机网络可以利用的各种信息资源的总和。网络信息资源是指以数字形式记录的，以多媒体形式表达的，存储在网络和计算机磁介质、光介质以及各类通信介质上的，通过计算机网络通信方式进行传递的信息内容的集合。在理解网络信息资源这一概念时，需要注意的是网络信息资源并非包含所有“投放”到因特网上的信息，它只包含能满足人们信息需求的那一部分。

（二）网络信息资源的不同类型

根据不同的划分标准，网络信息资源可以划分出不同的类型。

按时效性可以分为：网上出版物、动态信息、联机馆藏书目数据库、国际联机数据库。

按其分布可分为：企业、公司站点资源，大学、科研院所站点资源，信息服务机构站点资源，行业机构站点资源。

按网络信息资源的层次可分为：知识信息、信息单元、文献、信息资源、信息系统。

按照人类信息交流的方式可分为：第一，正式出版信息。指

① 吴静 . 基于 Internet 网络文献信息检索研究 [J]. 图书馆，2007（1）：109–111，118.

受到一定的产权保护、信息质量可靠、利用率较高的知识性、分析性信息，如通过万维网，用户可以查询到的各种数据库、联机杂志和电子杂志、电子工具书、报纸、专利信息等。第二，半正式出版信息。指受到一定产权保护的，但没有纳入正式出版信息系统的信息，如从各种学术团体和教育机构、企业和商业部门、国际组织和政府机构、行业协会等单位的网址或主页上，可以查询到但从正式出版物系统中无法得到的“灰色”信息。第三，非正式出版信息。指流动性、随意性较强的，数量大，质量难以保证和控制的动态性信息，如电子邮件、电子会议等。

（三）网络信息资源的主要特点

网络信息资源是指以数字化形式记录的，以多媒体形式表达的，存贮在网络计算机磁介质、光介质以及各类通信介质上的，并通过计算机网络通信方式进行传递的信息内容的集合。与传统信息资源相比，网络信息资源在数量、结构、分布、传播范围、类型、载体形态、内涵、控制机制、传递手段等方面，都与传统的信息资源有显著的差异，呈现出许多新的特点。

1. 信息复杂多样。网络信息资源具有大数量、多类型、多媒体、非规范、跨时间、跨地域、跨行业、多语种等特点。举凡文章、数据、图形、声音和视频等均列入其中，信息资源管理的复杂性和多样性空前增加。

2. 信息时效性强。网上的信息更新快。在互联网上能够得到最新的资料，获得某个科研的最新动态，能够得到最近或者当日

的文献。

3. 信息内容广泛。网络信息资源以电子文件的形式分布在世界各地的数以百万计的计算机上，包括不同学科、不同领域、不同地区、不同语言的各种信息，它们组成了世界上最大的信息资源库。

4. 开放性和交互性强。信息具有互动性，网络信息资源的共享性促进了网络信息资源数据结构标准化的进一步发展，使各种资源之间的开放性和兼容性越来越强。传统载体信息与用户之间是一种单向的关系，用户处于被动接受的地位。而网络信息资源具有互动性，用户的主动性增强，他们既是网络信息资源的利用者，也是网络信息资源开发的主体。比如，科研工作者既可以不受限制地使用网络资源，也可以将自己的优秀成果放在网上，供网络用户使用，成为网络资源的建设者。

5. 存取方式多样。在网络环境下，信息存取方式与所获得的信息资源类型和范围密切相关，其存取方式主要有电子邮件型、电话型、揭示板型、图书馆型、书目型等。

6. 信息资源的不稳定、相对无序性。互联网是开放的，通过TCP/IP 将不同的网络链接起来，对网络信息资源的组织管理无统一的标准和规范。同时，信息的地址、链接及内容本身处于经常变动状态，信息资源的变更、消亡无法预测。

7. 信息质量良莠并存。信息发布具有很大的自由性和任意性，隐私型信息进入了公共信息传播渠道。由于缺乏必要的过滤、质

量控制和管理机制，网络信息繁杂、混乱，信息质量良莠不齐。不仅学术信息、商业信息、政府信息、个人信息混为一体，而且负面信息也得以扩散，引发了多方面的问题。

二、网络文献信息资源建设概述

网络信息资源是图书馆的虚拟馆藏，而非图书馆实实在在拥有的。相对文献信息资源建设而言，网络信息资源建设是图书馆面临的一项艰巨的任务，如何采集、组织、开发网络信息资源是图书馆文献信息资源建设的重要内容。网络信息资源既包括图书馆网上发布的数字化馆藏信息，又包括从网上获取的各类信息资源。网络信息资源建设包括两个方面：第一，将自身资源进行数字化并送上网络；第二，通过网络获取别人的资源为自己的用户服务。

现阶段网络文献信息资源建设的内容主要有以下几个方面：

（一）传统文献信息资源全文数字化的加强

传统文献信息资源的全文数字化，即本地文献信息资源的全文数字化，它是整个文献信息资源建设的基础。传统文献信息主要包括印刷型文献（如图书、期刊等）、缩微型文献（如所谓胶片）、磁介质文献（如磁带、录像带）等。根据传统文献的特点和文献信息全文数字化的要求，多途径逐步进行数字化建设。

作为图书馆，由于条件的限制，不可能也不必要对所有的馆藏进行数字化，必须选择一些具有特色的文献进行数字化转化。要针对现有馆藏基础、馆藏发展计划、地区性合作计划、馆藏特

点等因素，确定数字化建设的具体内容，将现有馆藏文献信息有重点、有计划地进行数字化转化。馆藏文献的数字化转化还要避免低水平的重复浪费，提倡走地区性合作开发的道路。文献信息资源数字化转换是图书馆文献数字化信息建设的前提。

实现文献资源数字化的方式有以下几种：

1. 描录方式。将文献按原貌逐页存储为图像文件，并为其编制题名、作者、分类、主题词或关键词的索引，通过此索引可查找所需文献并显示其所在页面的图像文件。此方式可以保持文献的原貌，制作技术相对简单，减少了文件的数据量，便于该文件的整理、存储、传递和编目。

2. 文件方式。以文件方式存储文献内容，辅之以全文检索系统构成的全文检索数据库。全文检索是指数据库的全部字词编制索引，既可以采取逐字遍历的方式从文库中查找相应的字词，同时也可以配以题名、作者等索引进行辅助检索，或对文库中的人名、地名、关键词等编制规范文档，进行后控，以提高查全率和查准率。

3. 挂接方式。全文版挂接扫描版，即将上述两种存储方式结合起来。理想的制作方法是先制作扫描版，然后利用光学汉字识别技术将其转换为文本格式，从而建立文本与页面的一一对应关系。使用时，用户可用全文版检索到特定词所在的句段，必要时可调阅该句所对应的页面文件（扫描版）用以观看文献的原貌。

（二）馆藏电子文献采购的加强

电子出版物将是未来图书馆优先收藏的对象。由于数字化建设的经费有限，无法投入大量的人力、物力、财力进行所有馆藏资源的数字化，应有计划地采购现成的数字化馆藏电子文献资源产品，如《中国学术期刊（光盘版）》《中国学术会议论文》《中国学位论文》等，使馆藏电子文献所占比例逐步增加，这是建设数字化信息资源的有效途径。采购电子出版物时，不仅要注重各类检索性数据库的购买，更要注重全文型电子图书和报刊的采购。①

图书馆要调整采购政策，明确电子文献的建设方向，逐年增加电子文献的专项费用，还可通过地区性联合来采购大型的电子文献数据库，丰富馆藏资源，把有限的经费用在重要的地方。由于读者从印刷型文献阅读到电子阅读有一个过渡时期，而且电子文献的阅读对计算机和网络有依赖性，所以对于利用率高的文献，要考虑同时购置印刷版和电子版。

（三）网络获取信息资源的加强

互联网现已成为全世界最大的信息资源宝库，网络为人们展示了无限的空间和无限的资源，它所包含的信息异常丰富，从中可以挖掘出许多有价值的文化、商业、科技、经济等信息。开发和利用网络信息，延伸自己的馆藏资源，为用户实现“虚拟图书馆、无墙图书馆、全球图书馆”的高度资源共享，是数字图书馆建设

① 董素音，冯宝秀 . 现代图书馆藏书建设 [M]. 北京：海洋出版社，2006.

的主要任务。

由于互联网上的信息缺乏统一的组织管理和统一的目录，且信息资源是分布在整个网络中的，所以，人们在检索这个世界上最庞大的网络化信息资源宝库时常常感到无所适从。目前虽然互联网上已有谷歌、雅虎等比较有效的搜索引擎，用户可以利用这些搜索工具寻找一些资源，但是受时间和接触面限制，用户对网上资源了解不全。图书馆应该发挥职业优势，有重点地整理网络资源，并收藏到图书馆应用数据库中。目前，获取网上信息资源的途径主要有购买、租用和免费取得。获取网上信息资源的方法主要有以下几种：

1. 精心筛选信息。由于网络至今处于无序状态，面对网上各色各样的信息图书馆，应有选择地利用，要选择对服务对象和执行自身任务有用的内容和健康的信息。

2. 利用搜索引擎。虽然网上信息没有进行全面组织，也没有提供题录、文摘、索引等服务，但是除了电子邮件外，均有相应的检索工具及搜索引擎。这些检索工具可以分为检索性和目录性两种。检索性工具允许用户选择检索词查找信息，适用于专题检索；目录性检索工具则要求用户按检索系统事先设定的检索路径进行查寻，通过超文本链接按层次搜寻，它适用于宽泛检索。

3. 网上直接取用信息。在因特网上快速浏览时，会不断遇到一些有用的信息，可以直接将这些信息转移到指定的文档中以便利用。