

基于语义的 馆藏资源深度聚合与 可视化展示理论与方法研究

夏立新 翟姗姗 程秀峰 伍革新 著



科学出版社

国家社会科学基金资助项目

基于语义的 馆藏资源深度聚合与 可视化展示理论与方法研究

夏立新 翟姗姗 程秀峰 伍革新 著

科学出版社
北京

内 容 简 介

本书在理解、梳理语义网、馆藏资源聚合相关理论、方法与技术的基础上,以用户需求为导向,采用用户行为、本体、关联数据等理论思想,以及相关方法工具,以探究馆藏的知识内容以及知识单元的语义关系,通过关联、聚类、融合和分类,使得散乱和独立的资源再次组织成为一个有序整体,用以服务于用户对馆藏中所需的规律、模式、特色等知识单元以及知识之间的关联关系,最终实现基于语义的资源聚合和可视化展示平台。

本书可作为从事图书馆用户信息行为研究、信息构建、知识组织和实践的人员的参考用书;亦可作为情报学、图书馆学、档案学、计算机应用等专业的本科生、研究生教材。

图书在版编目(CIP)数据

基于语义的馆藏资源深度聚合与可视化展示理论与方法研究 / 夏立新等著.
—北京:科学出版社,2017.9

ISBN 978-7-03-052743-1

I ①基… II ①夏… III. ①数字图书馆-馆藏管理-研究 IV. G250.76

中国版本图书馆 CIP 数据核字 (2017) 第 101267 号

责任编辑:李 敏 / 责任校对:彭 涛

责任印制:张 倩 / 封面设计:李姗姗

科学出版社 出版

北京东黄城根北街 16 号

邮政编码:100717

<http://www.sciencep.com>

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2017 年 9 月第 一 版 开本:787×1092 1/16

2017 年 9 月第一次印刷 印张:17 1/4

字数:400 000

定价:120.00 元

(如有印装质量问题,我社负责调换)

前 言

2015 年 11 月,《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》发布,继续强调了国家文化事业发展的重要战略意义,指出:“十三五”期间要继续展开文化体制改革,建设重大文化工程,完善公共文化的产业、市场和服务体系发展。而图书馆作为文化传播和保障公民文化需求的重镇,应继续发挥其优势,并不断发展和变革,为广大人民群众提供更优质的服务。

进入互联网时代,图书馆资源由传统的实体走向数字化和网络化,用户对信息的需求也由原来的信息查找转向对知识的获取与个性化服务。历年来,各地、各单位图书馆的建设通过不断扩大其规模来满足读者的信息需求,但却造成“馆藏资源孤岛”和“馆藏资源超载”两大困境。而用户的需求不再仅仅是从无数馆藏中快速找到所要信息,还要求获得信息中蕴含的知识内容及逻辑关系,以进行知识的理解、使用和创造。针对图书馆的这种困境和数字化资源环境中用户更高的阅读需求等问题,通过多种途径、多种方法来揭示和关联日益丰富的数字资源,以利于用户对信息资源的有效获取和知识内容的共建共享。所以,如何对馆藏资源进行深层次揭示、聚合和有效地展示、利用成为业界关注的热点问题。

基于此,本书从理论、方法和实践角度探讨了如何从具有丰富语义信息的馆藏中进行资源的语义聚合和可视化展示,以期能够通过资源聚合更好地服务于用户。只有在馆藏资源的语义化基础上,才能通过共通的语义形式从多个方面对资源进行关联,进而实现从资源聚合到知识的创新。本书以用户需求为导向,采用用户行为、本体、关联数据等理论思想,以及相关方法工具,以探究馆藏的知识内容以及知识单元的语义关系,通过关联、聚类、融合和分类等过程,使得散乱和独立的资源再次组织成为一个有序整体,用以服务于用户对馆藏中所需的规律、模式、特色等知识单元以及知识之间的关联关系,最终实现基于语义的资源聚合和可视化展示平台。

全书共分为 10 章,每章的主要内容简述如下。

第 1 章,首先,从全局概述开展数字图书馆馆藏资源语义聚合的背景和意义,并从数字资源聚合、数字资源聚合的方式、数字资源聚合技术和数字资源聚合的实践探索四个角

度对当前馆藏资源语义聚合的基本现状进行梳理和述评。其次,根据现有馆藏资源的聚合方式及存在问题,总结得出在关联数据基础上的资源聚合的优缺点,其后提出完善馆藏资源的服务内容的研究思路和方法。最后,对本书的主要内容及创新之处进行了总结。

第2章,分析了当前数字资源聚合的基本概念和方法,根据数字图书馆中资源聚合出现的问题,归纳了当前数字资源聚合所面临的困境。在用户行为相关研究的基础上,阐述了国内外学者对用户行为的不同认识,给出了本书对用户行为的理解。进一步分析基于用户行为的数字资源语义聚合理论框架,具体从用户行为定义、用户行为理论和方法、用户行为数据聚合三个方面做了详细阐述。

第3章,论述了基于语义的馆藏资源语义聚合的相关理论。首先,对语义网及语义标注相关理论进行了阐述,主要包括语义网的发展、语义标注相关技术,并厘清了元数据、本体及语义标注间的关系。其次,介绍了关联数据的概念、关联数据的类型划分、关联数据的支撑技术及其构建的关键实现技术。最后,分析了基于关联标签的数字资源语义聚合方法的内涵、特点、步骤和优势。

第4章,深入对如何构建用户行为特征库进行了研究。首先,从构建用户行为特征库的外部系统环境入手,对用户行为的获取与挖掘进行了宏观研究。其次,从数据库的层次结构视角对构建用户行为特征库进行了可行性分析。接着从数据预处理、数据采集、数据抽取,到数据格式标准化对用户行为创建流程进行了详细论述,并对特征库的层次结构进行描述。最后,在本章实验部分,依照特征库创建流程设计了三个用户行为特征采集与挖掘工具。

第5章,基于用户行为特征库内的词语集合,创建用户行为词表,研究词词关联和词表聚合。首先,利用成熟的词表技术将用户行为特征库中的词汇进行序化,并借助 TF-IDF 改进算法及 MDS 算法两种方式来提取能代表用户特征的中心词汇。其次,讨论了模糊集合理论中词词关系的数据处理问题,在此基础上进行词表中词词关联值的探讨,并通过关联矩阵进行词间语义关系的实证研究。然后,借助比较排序、相关排序、Pagerank 等方法对词表进行融合研究。最后根据词间语义关系进行由词表向本体的转化,并阐述了如何利用词表创建用户行为的本体过程。

第6章,主要在第4和5章的技术基础上构建用户行为本体库,并以用户为中心进行本体聚合。首先,用户需求、用户体验、用户兴趣三者为用户行为本体构建的出发点,在本体创建总方向上对用户行为本体构建做出规范。其次,进行了用户本体的存储与操作的相关内容,在这些理论与方法的基础上阐述了用户行为本体的构建流程。再次,提出以用户行为特征为中心的异质本体聚合与融合的解决方案。最后,以湖北省图书馆的部分用户

行为数据为数据源,设计了“基于学科概念的用户行为本体聚合”实验。

第7章,基于已有学科概念分类,提出对用户行为主题概念的分类与聚合。首先,介绍了用户行为概念相关知识,指出概念聚合对于可视化导航尤为重要。其次,讨论了如何将聚合状态的主题概念体系以主题图的形式呈现出来,并对用户行为概念的主题图聚合进行了探讨。最后,在实验部分,通过对图书情报领域的个人博客为数据源,提出了利用主题图技术对以用户查询关键词为中心的,图书馆博客相关概念的链接与聚合方案,并将概念聚合形态进行可视化呈现。

第8章,从关联数据的视角对馆藏资源展开聚合研究。首先,阐述了关联数据的采集与处理方式,具体对馆藏资源聚合的必要性做了讨论,并介绍了数字图书馆关联数据的三种采集方式——被动服务器采集、数据端采集和客户端嵌入采集。其次,在分析了基于关联数据的数字资源聚合扩展和聚合模式的基础上,设计了基于云服务和数据仓库等技术的数字图书馆资源聚合及资源服务平台。最后,以 Drupal 关联数据平台为基础,对湖北省图书馆目录导航进行实证研究。

第9章,提出基于关联数据的数字图书馆创新服务。首先,从关联数据的知识发现、语义检索与知识服务3个角度探讨基于关联数据理论的知识发现与知识服务体系。其次,探讨采用语义方法的用户服务模型、信息推荐及服务集成,并由此对馆藏信息的个性制定与推送展开了探讨。最后,探究了以关联数据为方法的数字化虚拟参考咨询服务。

第10章,基于关联数据的非物质文化遗产资源聚合实证研究——以“楚剧”为例。本章选取了一类具有典型应用示范意义的馆藏数字资源——非物质文化遗产资源为应用对象,从基于语义的馆藏资源深度聚合与服务实践研究的角度,首先对楚剧资源聚合的必要性与可行性展开分析,构建楚剧资源元数据标准及领域本体,并对资源进行规范化标注。然后,对楚剧领域各类资源的关联数据进行创建与发布,实现资源语义关联。其次,搭建了非物质文化遗产资源聚合与服务平台,面向用户提供楚剧资源。最后,将本书构建的建立在关联数据之上的非物质文化遗产资源聚合与服务平台与传统信息系统进行比较分析从可行性与科学性两个方面展现出基于关联数据的非物质文化遗产资源聚合方法具有一定优势。

本书是作者及课题组多年研究成果的系统梳理与汇总,也是课题组成员集体智慧的结晶,许多同志对书中涉及的内容作出了贡献。本书由夏立新主笔统稿,夏立新负责第1章的写作,夏立新、翟姗姗负责第2、3、10章的写作,夏立新、程秀峰负责第4、5、6章的写作,夏立新、伍革新负责第7、8、9章的写作,书中的绘图部分由白阳、郑路完成,李重阳、康雨培、毕崇武、张玉晨参与了本书的文字校对工作。

本书的研究得到了国家社会科学基金重点项目（11AZD090）及中央高校基本科研业务费（人文社科类）重大研究领域培育计划：知识管理与服务（项目编号：CCNU14Z02002）的资助，科学出版社对本书的出版给予了大力支持，在此一并致谢！

由于作者水平有限，加之时间仓促，书中可能还存在一些不足之处，恳请各位读者批评指正。

夏立新

2017年3月20日于桂子山

目 录

前言

第1章 绪论	1
1.1 问题的提出	1
1.2 馆藏资源深度聚合与可视化展示的研究价值	2
1.3 国内外数字资源聚合研究工作综述	4
1.4 本书的主要内容	12
1.5 本书的创新之处	13
第2章 面向用户的数字图书馆资源聚合相关理论	15
2.1 数字图书馆数字资源聚合有关理论	15
2.2 用户行为研究理论	27
2.3 基于用户行为的数字图书馆资源聚合理论框架	29
第3章 基于语义的馆藏资源语义聚合理论基础	32
3.1 语义网及语义标注相关理论	32
3.2 关联数据有关理论	36
3.3 基于关联数据的数字资源语义聚合方法	48
第4章 用户行为特征库的创建	56
4.1 用户行为特征库的系统环境	56
4.2 用户行为特征库的可行性分析	65
4.3 用户行为特征库的创建流程	67
4.4 用户行为特征库的结构与描述	74
4.5 实验：用户行为数据挖掘工具的开发	78
第5章 用户行为特征的关联与序化	84
5.1 用户行为词表	84
5.2 用户行为词表的自动创建	91
5.3 词词关联矩阵及词间关联值的算法	93
5.4 实证：词表关联矩阵词间语义关系验证	98

5.5	用户词表与现有词表的融合	103
5.6	用户行为词表转化为用户行为本体	110
第6章	基于本体的用户行为数据语义聚合	114
6.1	用户行为本体构建的出发点	114
6.2	本体的存储与操作	119
6.3	本体的构建流程	129
6.4	本体的语义聚合	137
6.5	实验：基于学科概念的湖北省图书馆用户兴趣本体聚合	141
第7章	基于主题概念的数字资源聚合与可视化	148
7.1	用户行为主题概念聚合理论	148
7.2	基于概念聚合体系进行主题图创建	161
7.3	以用户行为概念为中心的主题图聚合	169
7.4	实验：基于主题图的图书馆用户个人博客主题概念聚合	173
第8章	基于关联数据的数字图书馆资源聚合	179
8.1	关联数据的采集与处理	179
8.2	基于关联数据的数字图书馆数字资源聚合内涵	181
8.3	基于 Drupal 的关联数据平台	188
第9章	基于关联数据的数字图书馆创新服务	197
9.1	基于关联数据的知识发现与知识服务	197
9.2	基于关联数据的信息个性定制与推送	205
9.3	基于关联数据的数字化虚拟参考咨询	214
第10章	基于关联数据的馆藏资源语义聚合实证研究——以非物质文化遗产中 “楚剧”为例	221
10.1	“楚剧”资源聚合的必要性与可行性分析	221
10.2	“楚剧”资源采集及语义描述	223
10.3	基于“楚剧”资源的关联数据创建与发布	234
10.4	非物质文化遗产资源聚合与服务平台应用实例——“楚剧”资源展示	240
10.5	“聚合平台”与传统信息系统的对比分析	252
参考文献		254

第 1 章 | 绪 论

图书馆数字资源建设正在全国如火如荼地展开，各大图书馆也增强了对用户需求分析、用户行为分析的能力，以期使数字信息服务更加贴近用户需求（程秀峰，2013）。与此同时，以知识检索、知识发现、知识利用和知识服务为根本目标的数据组织与整合一直是数字图书馆研究的重点。

Web3.0 网络环境下，知识共享进一步深化，各部门信息化工作进一步深入。在图书馆领域，一场将知识网络化、数字化、语义化的变革促成了数字图书馆的快速发展。习近平主席多次在不同重要场合强调，要建立社会主义文化强国，不断丰富人民精神世界，着力提高文化软实力。不言而喻，在知识经济大潮下，利用先进信息技术，对国家丰富的图书馆馆藏资源进行现代化的保护和处理，通过语义网、资源聚合等先进技术使图书馆数字资源焕发出新的生机，这对促进国家文化创新，增强文化发展活力具有重大的现实意义和理论指导意义。

在绪论部分，笔者首先从整体上概述了展开数字图书馆馆藏资源语义聚合的背景和意义，以及当前馆藏资源语义聚合的基本现状，并从整体上介绍了本书的研究思路和方法。

1.1 问题的提出

进入 21 世纪以来，国家数字图书馆在网络化、数字化、自动化方面成绩斐然。早在 2001 年，全国已有 87% 的图书馆实现了上网功能，71% 的图书馆开设了特色数据库，其中有超过半数的图书馆参加了区域性和系统内的信息资源共建活动（高波和刘兹恒，2001）。绝大部分高校图书馆实现了从传统的纸质文献为主的采集策略向以开发数字资源为主的策略转变，并不断提升知识服务功能。

然而，长期以来，图书馆信息平台在操作系统、数据结构、数据存储、网络传输上存在严重的异构性。这种异构性导致用户访问数字资源时除了需要克服知识产权、知识保护等障碍外，还需要克服资源分散、重复查找等困难，这大大降低了用户使用数字图书馆的兴趣。相关文献资料表明，在网络化、数字化时代，用户对数字图书馆的利用率并不高。目前，图书馆数字资源需求最大的群体主要是科研工作者、高校教师和学生，而 OCLC 对全国大学生图书馆和信息资源认识的报告显示：绝大多数大学生用搜索引擎进行检索，只有极少数的人从图书馆进行信息检索，而且绝大多数人满足于搜索引擎的体验大部分人对搜索引擎的效果表示满意（贺兰芳，2010）。相比而言，84% 的大学生对有图书馆员帮助的搜索经历表示非常满意或满意（常唯，2006）。可见，国家数字图书馆建设工作不仅需要大力加快知识服务、知识更新与知识链接，使蕴含在馆藏资源中的知识易于获取，更需

要以用户需求为导向,立足于知识集成与知识聚合,增加数字资源的利用率。另外,数字图书馆对于用户的需求、体验与兴趣方面还主要停留在需求调研、问卷调查、兴趣讨论的阶段。如何定量分析用户行为,利用词表、本体、主题图等技术方法将用户行为的客观规律与数字资源的分配相结合,将用户行为融入异构数据组织与整合中,并最终建立服务于语义检索和语义导航的资源聚合体系,这是图书馆进行数字资源科学组织的关键。

国外许多数字图书馆在分析用户行为方面做了大量工作。这些工作从用户行为预处理、用户行为模式的发现、用户行为模式分析3个层面对数字图书馆用户的存在形式、存在作用、存在范式、兴趣偏好等进行了深入研究,形成了以数据挖掘技术为支撑的用户行为研究模式。这些研究模式从手段上来说,均是以数据异构性为切入点,在语词关联、概念聚类、数据互操作的基础上对大量纷繁复杂的用户数据进行重新布局,并且从用户的体验与实际需求出发,强调系统与用户的认知协调,将用户行为反馈融入信息检索机制中,强调以用户需求为导向的数字资源科学分配(晁亚男等,2012)。在系统开发方面,以美国为例,既开发出了以OPAC为聚合手段的联机索引机制,还发展出一套基于元数据的开放式链接标准。另外,OCLC、国会图书馆(2016)、MELVYL联合目录(2016)等一系列不同类型的资源整合系统,更是将美国图书馆资源极大程度上聚合起来,为用户的查、阅、借提供了统一且便捷的服务措施。但是,对于数字图书馆用户数据的挖掘并形成对馆藏资源深度链接与聚合的相关研究,目前并未受到相应的重视。

因此,要在图书馆资源建设中融入用户行为数据,并最终实现语义化的信息检索与信息利用,就要在数字图书馆资源整合的大背景下,利用先进的资源整合与链接工具对图书馆数字资源进行统一描述、映射、链接与聚合,并根据用户需要,把分散的、异构的数字资源集中,形成一套整体的聚合体系标准,从而达到一个知识共享与发现的新层面。

1.2 馆藏资源深度聚合与可视化展示的研究价值

如何解决数字图书馆资源间的语义不对称问题,提升数字资源的利用效率,促进知识的交流和共享,实现图书馆资源的无障碍化利用,研究者提出了多种解决办法。其中比较有效的方法是借助语义技术进行馆藏资源的聚合,数字资源对象经过语义方法发布之后,具有可共享、规范化等特点,便于对数量众多、类型复杂的馆藏资源展开聚合,构建相同及不同学科的资源间的链接,进而建立跨库检索,实现对数字图书馆资源的聚合。在语义推理扩展的基础上,用户能够在检索时明确语义,并对语义索引的关联数据进行构建,从而对用户所需的数字资源进行检索,通过一系列如本体和子图等语义搜索,对进一步发现更多用户所要的资源以及资源之间的关联有一定的帮助,从深度和广度两个方面加深数字图书馆的资源服务。

1.2.1 实践价值

以语义为基础的馆藏数字资源聚合与可视化呈现是针对数字资源无序且难以有效获取

和利用的问题提出的。以关联数据的技术构建为支撑,对数字图书馆数字资源进行统一标注,一方面有助于迅速查找与定位用户所需的跨库和跨平台资源,另一方面保障了知识内容的高质量。采用关联数据进行数字资源聚合的实践意义如下。

(1) 实现数字资源语义层面的知识聚合与服务

数字图书馆数字资源服务主要是能根据用户所需提供必要的知识服务。就经济效益而言,最基础的是资源产品的使用价值及用户需求,能够为用户提供较优质服务是数字图书馆价值所在,应谋求以多元化发展为标准,立足现实,长远兼顾。数字馆藏资源的有序化与服务密切联系,数字图书馆解决的关键问题是要加强数字资源之间的语义关联,以及清晰的可访问的途径。随着科技的发展和用户日益多元化发展的趋势,图书馆服务方式也随之从以文献为核心的服务逐渐转向以用户为中心的知识化服务。若要实现知识服务的发展,首先需要解决馆藏资源的统一标准和知识的细粒度划分问题,使数字馆藏资源能够实现语义关联。因为借助简单的图示无法清楚展示知识单元之间复杂多样的关系和结构,而且随着用户需求的提高,常用的关键词匹配和链接技术越来越无法满足需求,而通过关联数据等语义方法对资源实现聚合,可以帮助揭示资源内容结构和含义,满足用户对更细粒度的知识要求。采用关联数据实现数字资源的统一标识,对于数字资源聚合的多样性和复杂关系的简化,以及向用户呈现可动态变化的图形化知识脉络,具有一定的帮助,并可进一步提供知识导航和检索,促进用户知识服务。无论从学科体系结构及发展趋势,还是以用户为中心促进数字资源的知识服务水平的提升来看,对馆藏资源的语义化开发利用都具有理论和实践意义。

(2) 着力解决多维度数字资源深度聚合的问题

为解决数字资源聚合与用户知识服务的复杂系统工程,一方面需要考虑理论和方法论的问题,另一方面对于可操作方案须进行相应的探讨。根据用户不断提高的需求,数字图书馆的服务形式应从资源为中心的文献服务转向用户为主导的知识服务,结合数字资源建设与用户体验,以统一部署与因地制宜相适应为原则,构建基于语义的数字资源聚合平台。要实现平台的构建,既要深度揭示复杂、多样、迥异的数字资源,挖掘知识单元及关联关系,实现数字资源的高度聚合(孙丽霞,2010),又要实现对知识单元及其关联的可视化方式呈现,以简单的导航展现复杂的知识逻辑,保障用户知识服务的高质量。在这些功能实现的基础之上,国家文化资源的深度开发和利用将会得到进一步快速发展。

1.2.2 理论价值

本书对采用语义关系来改善数字馆藏的内部关联聚合、完善知识服务体系并研究适行的聚合方法、加强数字资源管理和关联学科领域的发展都具有重要的理论支撑作用。

(1) 深化和拓展数字馆藏聚合并完善知识服务体系理论的探究

本书主要在理论层面回答如下几个核心问题:如今数字馆藏聚合进程中会遇到什么样的问题,能否获取更优化的语义方法以优化数字资源的聚合;新的语义方法对数字馆藏聚

合提供了哪些支持；如何构建数字图书馆创新与知识服务体系，以及对关联数据技术作为支撑的数字资源聚合带来何种影响。就以上几个问题，本书将从如下几个方面进行探究：数字馆藏聚合的新方法——关联数据；在关联数据方法的基础上，实现数字图书馆资源聚合的详细过程；在图书馆资源聚合基础上，探讨数字图书馆知识服务的内涵与外延。最后构建以用户为中心并具有多维度关联特性的数字图书馆知识服务体系，该服务体系具有多维关联特性。

(2) 梳理基于语义的数字资源深度聚合的方法体系

数字馆藏聚合的实践活动中，无论何种类型、何种层次的聚合都必须依赖专属的方法与技术手段。通过使用语义方法，对基于元数据的数字化数字资源整合时，利用统一资源定位符（uniform resource locator, URI）实现对相关数据集的统一标识，并利用资源描述框架（resource description framework, RDF）三元组对数据集进行基于元数据的数据关联，从而有效整合数字图书馆数字资源（赵梓彤和谢海先，2011）。同时，分布式构建技术、互操作技术、知识处理技术、新型的知识组织技术、语义 Web 技术等，也是数字馆藏聚合实践中可能用到的重要技术手段（张寅生，2007）。实现语义层次上的数字资源聚合，需依赖于新的聚合技术与方法，以推动数字资源聚合研究向更深、更高的层面发展。以语义为导向的数字资源聚合的理论研究及实践应用得益于其方法体系的建立，方法与技术的创新，在理论思路以及材料方面为基于语义的数字资源聚合提供支持，有利于将数字资源聚合应用于实践之中。

(3) 促进数字资源管理、知识服务的发展

形成于 20 世纪 80 年代前后的数字资源管理是综合性学科，并没有因为历史短而限制其发展迅猛的势头，数字资源管理的研究仍在不断进行和深化。数字图书馆资源，作为数字资源重要组成部分，尽管和其他类型的数字资源有交叉和重叠的部分，但在本质上具有独立性。因此将图书馆数字资源作为独立的数字资源类型，将其作为研究对象全面深入地分析数字馆藏的内涵、功能及分布等，并进行基于关联数据的聚合研究，对于数字资源的整合、检索、利用与管理具有重要的应用价值。基于语义的数字馆藏聚合，能给用户带来检索上的便利，能提供一站式检索，并满足用户的细粒度化检索；从内容上可将数字馆藏进行语义关联、语义指向和实现信息的网状结构化，提高检索的查全率与查准率。本课题的研究，一方面能促进数字馆藏资源建设、数字资源管理和用户知识服务的发展，另一方面也将对一些学科的发展产生积极的影响（田依林等，2011）。

1.3 国内外数字资源聚合研究工作综述

1.3.1 国外数字资源聚合研究进展

国外学者研究始于 20 世纪 90 年代，对数字资源聚合、语义网、关联数据等方面进行广泛的研究，且研究成果颇丰。近些年，国外在语义数字图书馆等方面的研究取得了一

系列具有应用意义的理论和应用成果。

1.3.1.1 数字资源聚合

早在1967年,美国成立的联机计算机图书馆中心 OCLC,是世界上最大图书情报服务机构之一,其数字资源服务不仅针对国内数字资源,更将数字资源共享的范围扩展到了世界范围。2003年第五届世界图书馆联盟欧洲会议讨论了包括联盟建立、成本分配、评估在内的联盟管理问题;Elsevier Science 的“Big Deals”和发展的合同商议问题;以及开放存取问题,特别是 Budapest 开放存取创世计划和 SPARC Europe 项目(Yamamoto et al., 2004)。2011年举办的第三次世界图书馆联盟欧洲会议,也对电子资源的采购问题进行了探讨(Holmstroem, 2012)。

近些年,国外一些学者开始关注于复合图书馆的研究。Pinfield (2016)提出了复合图书馆的“无缝整合”,其原理是在统一的界面上为用户提供多样化的服务,并且认证的网关也是唯一的。同时 S. Pinfield 还提出了其内容与资源的互操作问题。数字图书馆数字资源和数字资源格式的多样性是资源互操作中的障碍。要实现数字资源的互操作就是要解决图书馆数字资源中的元数据互操作的技术层面上的难题。作为图书馆的未来发展趋势的复合式图书馆,其主要包含3个层次的内容(黄晓斌等, 2005):信息地图、代理结构以及数字资源供给系统。而基于语义的开放关联框架能够在判断和指向数字资源间的相互关联关系的基础上为用户提供多维度的数字资源服务,并有效促进复合式图书馆的建设与完善(An et al., 2009)。

同时,学者们还提出了数字图书馆联盟的数字资源整合方式。Chen Xiaotian (2002)对比了伊利诺伊州的 ILC SO 和密苏里州的 MOBIUS 两个学术图书馆联盟的特点,包括联盟政策对用户的影响、OPAC 的特征及文献的传递等。Shachaf (2003)对全国性图书馆联盟声明周期问题展开了研究,绘制出了图书馆联盟从萌芽、早期发展到成熟的生命周期和发展模型。Friend (2002)认为数字图书馆联盟作为数字图书馆发展的重要阶段,其功能不是简单的馆际互借和文献传递的合作,更需要关注期刊的采购,以保证更低的价格和更有利的许可条款。而 Dorner (2001)研究了在降低成本的同时改进数字图书信息服务的传递,这也是新西兰数字图书馆共同采用的一种新的在线图书馆联盟模式。

1.3.1.2 数字资源聚合方式

本书根据数字资源聚合层级的不同,将国外图书馆数字资源聚合方式归纳如下。

(1) 以数字资源内容为对象

以数字资源内容为对象的聚合,主要是将不同格式、不同存储形式、不同类别的数据进行信息聚合,也是数字图书馆数字资源聚合的前提条件。按照分类方式的不同,可以将当前图书馆的数字资源分为如下几种:①按数字资源的表现形式可分为文本、图形图像、视音频等。②按载体形式可以分为印刷型、数字型、虚拟环境型以及网络多媒体型等。③按照资源的组合形式,有结构化的与非结构化的数字资源。④按数字资源的地域分布,可分为本地数字资源和远程数字资源。⑤按资源的组合程度,可以将其分为非集成的分布

式数字资源和集成资源。

(2) 以提供数字资源全方位服务为目标

随着大众对图书馆数字资源聚合的需求不断深化,数字资源聚合的行业领域不断增多,内容范围也在逐渐扩大。数字图书馆数字资源聚合除了自身数字资源聚合以外,其关注点越来越偏向于为用户提供多方位的信息服务。在当前的数字资源聚合过程中,由于资源的分类导致数字资源的分布式存储和数据库管理的分类与存储方式的差异,从而在某种程度上降低了资源聚合的有效性和时效性。所以,本书提出的数字资源聚合正是着重解决分类和存储差异性问题和实现多学科、多领域数字资源间的服务共享,其中最为适宜的是一站式服务方案。

(3) 以语义网环境下数字资源聚合技术实现为对象

现行的图书馆多是通过物理存储介质存储资源,还有一部分通过数字化的存储形式来解决数字资源的存储问题,这种不同的存储方式的产生是因为忽视了资源内容本身的语义关联。而在 Web 的基础上建设语义 Web 的蓝图很早就被互联网的创始人 Tim Berners-Lee (2001) 提出,即网络中的各类型的资源能够在语义含义的层面上进行清晰的关联和划分,让机器能自动进行识别、分析、解释和执行。

语义网的出现,为数字图书馆建设带来了方便,国外学者致力于将语义网应用到数字图书馆建设中,为数字资源的语义揭示及深度聚合提供技术支持,最终实现知识整合与发现。这方面研究主要有构建以语义为基础的信息书目检索系统、利用本体构建数字图书馆等。

实现数字图书馆数字资源语义聚合是以本体为核心思想和方法,研究者对基于本体的数字图书馆数字资源的语义聚合进行了深入探讨。如 Bloehdorn 等 (2007) 探讨了数字图书馆中使用本体解析知识内涵并在此基础上提供知识查询,认为本体技术作为更有效的技术促进了整合不同来源的文献,在对知识库中的各类语义文档的标引、归类和划分的基础之上,为用户提供本体语义元数据的分类导航和语义检索查询等高级知识服务,将语义网应用于信息书目检索系统中的研究, Martin Malmsten (2008) 对 LIBRIS 符合语义万维网和 linked data 协议规范的工具和技术展开了研究,对精确描述个体资源的问题进行了讨论,对资源之间的链接以及数据获取机制等方面的内容进行了重点介绍; Christopher Yang 等 (2008) 则设计并构建了一个跨语言自动的语义词库,为数字图书馆实现跨语言语义分析和检索奠定了实践基础。

1.3.1.3 数字资源聚合技术

多种信息技术的采用是实现图书馆数字资源聚合的关键因素,由于数字资源为不同类型,需要借助和采用不同的聚合技术来实现数字资源的聚合。Roller (2002) 认为数字资源的聚合体现在数字资源的聚集与所提供服务功能的整合两方面,因此其相应的聚合技术即为如下两种。

(1) 数字图书馆数字资源聚集技术

数字资源描述技术是指从原有的数字资源库中根据元数据的分布不同和类别划分,对

其建立相应的语言标注模型,在数据模型方面为信息整合提供支持 (Malone et al., 2004);信息呈现技术是指对数字资源的知识内容进行组织和处理,依托知识的关联关系和内部逻辑,将数字资源的整体脉络呈现给用户,实现用户检索过程和检索结果的可视化。数据信息的抽取是使抽取的信息内容具有稳定的数据化结构,并将这些数据化结构提供给用户进行查阅,而这些抽取的信息内容则是按照资源内容和一定的划分标准、用户需求来进行抽取。这种以结构化的数据呈现形式能够让用户对数据资源中的知识内容进行描述,在一定程度上有利于实现用户的高效查询检索 (闫志红, 2008)。

(2) 聚合技术与其他技术融合

诸多聚合技术中,功能聚合技术主要有适配器技术、中间件技术、封装器/协调器技术等。适配器技术主要是转换不同应用系统、平台中包含的中间数据,统一数据源的接口,并能促进其他技术的融合;中间层数据转换技术是对结构或内容不同的数据源进行相互链接,实现数字图书馆各个结构的数据源与系统之间无缝连接;封装器是整合不同数据集中的数据资源,将它们统一为同种格式;而协调器技术则是有效融合数字资源整合中的用户信息、各数据源资源信息,以及系统应用之间的协调信息。此外,P2P、网格等技术被广泛应用于对异构系统功能与服务的整合 (Leymann et al., 2002)。

国外非常关注应用语义网来挖掘某类文献的语义信息,语义网使得数字资源系统功能与服务推向了一个新的台阶。如 Shahrul 等 (2010) 对学术文献中语义分类和语义管理等内容进行了详细的分析,认为其中采用本体的检索方式更为高效。Andrea 等 (2011) 致力于构建一个特定的文化遗产管理平台,在构建存档和平台层面的同时对数据异构性进行了改进。该平台还引入了 tag 技术,用该技术对知识进行组织、描述并建立知识之间的链接或检索。不同类别的 tag 标签应用于相应的知识模块,可以构建基于 tag 标签的聚合。

1.3.2 国内数字资源聚合研究现状

近 10 年来,对数字资源深度开发利用的相关问题已成为国内图书情报领域关注的重大课题。为此,本书从资源整合的理论、方法与技术及实践应用等方面对研究现状进行了如下梳理。

1.3.2.1 数字资源聚合相关理论

当前信息与网络技术的迅猛发展,使得数字资源呈几何级数增长,而数字资源的管理却越来越依赖于信息管理系统的应用。然而海量数字资源所呈现出的分散、无序以及不确定性,现有的信息管理系统难以有效解决。由于对数字资源的组织、整理、揭示与优化的能力无法适应于数字资源的快速增长,所以在某种程度上数字资源整合制约着数字图书馆发展和服务。

(1) 数字资源聚合概念界定

聚合,是包括数字资源的聚集与整合两方面。聚集简单而言就是将资源进行集中,而

在网络环境下的聚集则是一种虚拟的集中,是对数字资源间进行相互关联和链接指向。整合是将数字资源有序化处理,即对数字资源内部各数据模块之间的层次和结构进行层次化和结构化处理,使该数据资源形成一个有效的整体。而数字资源整合的更高阶段是融合,即对数字资源进一步的有序化的过程。杜小勇和马文峰(2005)认为图书馆资源的“集成”可以称为“整合”,主要有两层含义:一是聚合,利用现代信息技术将异构的数字资源聚集到一起;二是融合,即在数字资源聚合的基础上通过技术手段对多样化数字资源进行优化、组合,从而方便数字资源的获取,实现资源的高效利用。数字资源整合又称为数字资源集成、数字资源聚集、数字资源融合等,是将不同数据来源、知识领域、数据类型的数字资源聚集到一起,再经过技术处理将其进行规范化、结构化,让其形成相互关联、内容聚合的资源整体。

(2) 数字资源聚合的理论与方法

马文峰(2002)曾提出系统论的方法来聚合数字资源,既是数字资源系统构建的基本理论基础,也是数字资源聚合的基本方法依托。通过系统论观点和方法应用到资源整合的方法论原则以及数字资源聚合具体方法之中,从而达到数字资源整体优化;苏新宁(2005)也强调,除了图书情报学、信息学自身的诸多理论作为理论支撑之外,在提高数字资源整合效率和服务效果方面还可以借鉴系统论、运筹学、心理学等理论。

当前国内有关数字资源聚合的理论与方法的研究较多,出现了很多有代表性的成果。马文峰(2002)对数字资源的聚集、整合及系统构建等基本原理进行了研究。马文峰(2003)同时分析了数字资源整合的方式,并详细探究了数据整合、信息整合以及知识整合这3种概念,通过分析比较呈现了各自的特点。

在数字馆藏的整合过程中对数据、信息和知识3类资源的整合是普遍关注的问题,三者之间存在逐级递进、交叉重合却又不可替代的关系。其中数据整合是基础,知识整合则是数字图书馆资源聚合的最高形式,未来图书馆数字资源优化组合将朝着知识整合的方向发展。黄晓斌等(2005)就数字图书馆资源整合的必要性与基本原则进行了分析,认为数字图书馆资源整合过程中应遵循多项基本原则,并提出了数字资源聚合的未来发展趋势及方向。李广建等(2007)就国外的数字资源整合研究进行了分析,将其研究概括为3个层次:面向资源组织的整合、面向异构数据的整合及面向应用的整合,还对数字资源整合系统的体系结构做了研究。

1.3.2.2 数字资源聚合的方式

数字资源聚合呈现出不同的层次和不同的方法,按照不同应用技术,不同的角度对数字资源聚合其进行划分,主要有如下几种。

(1) 异构异类数字资源

赵冬梅(2005)曾提出了8种具体的聚合形态,分别是:跨库检索系统的聚合;期刊与数字资源内容的聚合;电子资源链接地址的聚合;目录式内容链接的聚合;全文电子期刊和电子期刊论文索引内容、文摘地址的聚合;学科信息门户内容的聚合、学科导航和学科信息智能导航聚合、学科电子资源搜索系统。