大数据时代图书馆业务管理与服务创新

马启花
（广西教育学院图书馆，广西 南宁 530023）

摘 要： 文章简要概述大数据的来源、定义及特点，探讨图书馆与大数据之间关系以及图书馆在大数据时代业务管理与服务方面创新。

关键词： 大数据；图书馆；业务管理；服务创新

中图分类号： G252.4 **文献标识码：** A **文章编号：** 1002-1248（2014）11-0205-04

On the Service Innovation and Business Management of Library in the Big-data Era

MA Qi-hua
(library, Guangxi College of Education, Nanning 530023, China)

Abstract: This article brief provides the origin, definition and characteristics of the large data, and explores the relationship between the library and the large data. Lastly the article Puts forward some ways in the service innovation and business management of library in the Big-dataera.

Keywords: Big data; Library; Business Management; Service Innovation

随云计算、Web2.0 后，大数据成为时下 IT 行业最火的词汇，实际上大数据这个概念在 20 世纪 80 年代美国就有人提出过。那么大数据从何而来？大数据来源于人类、信息系统以及实物。我们人类每天使用的手机，已经不单纯是一种简单的通讯工具，还可以产生视频、音频、相片等数据。我们平时进行 QQ 聊天，有聊天记录数据，可以上传文件、图片与人分享。计算机系统、集成系统每天运行产生的日志，系统内数据备份、拷贝等都产生大量的数据。大自然实物自身产生大量数据。如运用于交通安全的摄影头，波音 787 飞机飞行任务等。随着网络技术、通讯技术的发展，数据递增速度达到前所未有。每天 Google 需处理的数据量约 24PB，百度约几十 PB，淘宝网约 20TB（1TB = 1,024 GB，1PB = 1,024 TB ，1EB = 1,024 PB，1ZB = 1,024 EB，1YB = 1,024 ZB）。根据国际数据公司 IDC 预测，到 2020 年，全球将拥有数据量达 35ZB。人们工作、生活包围在大数据周围，工作、生活方式将随大数据的到来发生根本性改变。

1 大数据的概念及其特点

关于大数据的概念，目前没有统一的界定。最早提出大数据概念的是麦肯锡，他认为：数据在当今已经渗透到每一个行业和业务职能领域，并将逐渐成为重要的生产因素。人们对海量数据的挖掘、分析及运用将预示着新一波生产率增长及消费者盈余浪潮的到来。但彬认为：“大数据”包含了“海量数据”的含义，而且在内容上超越了海量数据，简而言之，“大数据”是“海量数据”＋复杂类型的数据[1]。John Rauser 认为：大数据是任何超过了一台计算机处理能力的庞大数据量[2]。也有学者认为“大数据是指那些大小已经超出了传统意义上的尺度，一般的软件工具难以捕捉、存储、管理和分析的数据”[3]。尽管众多学者对大数据进行多个定义，没有形成统一，但也达成一种共识，即大数据之“大”不仅代表数据的海量，代表更多的属性。为此，IDC 总结了大数据具有的“4V”特性，即种类多(Variety)、流量大(Velocity)、容量大(Volume)和价值高(Value)[4]。

1.1 数据类型种类多

除传统的结构化数据(我们熟悉的文本数据)外，还包括非结构化数据和半结构化数据，如音频、网络日志、动漫、相片、视频、地理位置信息等。异构化数据呈持续增长，种类繁多，它们之间相互作用，相互

收稿日期： 2014-05-14

作者简介： 马启花（1970-），女，研究馆员，硕士，广西教育学院图书馆从事图书采访工作，发表论文 30 多篇，主持省级、厅级、院级重点等课题多项，研究方向：文献信息资源建设。

关联使数据局面更加复杂化，对数据处理分析能力提出了更高的要求。

1.2 数据流量大增长速度快

目前数据以呈指数级速度增长，储存单位从 TB 级别跃升到 PB 乃至 EB 级别。据调查，最大的数据仓库中的数据量每两年就增加 3 倍。按此速度发展，到 2015 年该数据库的数据量将逼近 100PB。大数据的快速发展彰显其实时动态性强，这就要求及时响应，分析处理速度要快，时效性要求高，否则有些数据稍纵即逝。

1.3 数据总量规模巨大

“大”是指大型数据集，超越常规管理、处理和分析，一般数据量在 10TB 以上，以后 PB 级别将成常态。目前一些大企业其数据量已达 PB 级别，全球数据总量已达 ZB 级别。IDC 报告显示，未来 10 年全球大数据将增加 50 倍。

1.4 数据价值密度较低

在海量的数据里，信息无所不在，价值无所不在，但密度相对较低，如何在庞大的数据中快速地发现并“提纯”到高价值的信息，就需要有强大的机器算法，这是大数据时代亟待解决的难题。

大数据在改变人们的生活方式，颠覆传统商业模式，成为一种战略资源，一种竞争要素。目前人们已意识到大数据价值所在，并开始挖掘、分析、发现并利用其价值。沃尔玛“啤酒 + 尿布”就是一个成功的商业案例。图书馆是信息聚散地，是收集、处理、分析、存储及利用信息的学术性服务机构，如何在大数据时代抓住机遇提升自身竞争能力，如何进行自身数据管理，如何在用户众多非结构化、半结构数据发现用户阅读习惯及阅读规律，并应用到文献信息资源建设，从而提高馆藏质量，这些都是图书馆人应该认真思考的命题。

2 大数据与图书馆

2.1 图书馆自身具有大数据特征

图书馆是个生长有机体，随着周围环境变化发展而发展。一项新技术的出现，图书馆都会自觉利用新技术来发展自己，壮大自己。计算机技术、网络技术的出现，致使图书馆业务集成管理系统面世并投入使用，提高了图书馆业务管理能力和服务效率，也产生了业务数据、工作日志。电子刊物，电子资源也如雨后春笋般发展起来，网络资源，纸质图书回朔建库，图书馆创建特色数据库等都产生大量数据。图书馆 Web2.0 出现，图书馆利用 Web2.0 技术，如微博，维基等与用户交流，也产生了大量的数据。数字化建设，数字图书馆产生了大量数据资源。随着智能手机、移动图书馆的普及，数据量将呈指数上升趋势。网络化、信息化技术的发展使图书馆初具大数据特征。

2.2 图书馆提升业务管理与服务水平需要大数据的支持

图书馆业务管理及服务水平的提升受方方面面的制约，如图书馆集成管理系统、出版业发展水平、供应商综合实力情况、馆藏质量、读者阅读行为习惯等。这些因素的信息是零散的，有些甚至是不对称的。在大数据时代，利用大数据成熟的技术，把来自上述各因素零碎的庞大数据整合在一起，可以构建出图书馆业务管理和服务的全景图。洞察各因素的细微变化，从而做出快速响应，制定有效发展策略。庞大的相关的数据整合在一起，就能不断地产生新的信息和知识，有助于提高图书馆生产效率、降低经营成本。且在大数据时代，图书馆间的竞争不仅仅是看你拥有馆藏资源多少，建筑空间多少、服务水平怎样，而是看你拥有大数据量是多少，是否拥有挖掘、分析处理大数据的技术。大数据将是图书馆未来核心资产。

2.3 大数据对图书馆的影响。

大数据有利提升图书馆核心竞争力，但也对图书馆提出了挑战。（1）数据收集问题。在大数据时代，图书馆要全面地准确掌握信息，才能作出正确决策，这就要求图书馆既要收集图书馆内部各业务流程数据，又要收集图书馆外部的相关数据，需要巨大存储空间，需要耗费大量的人力、物力和财力。（2）图书馆集成管理系统软件问题。大量数据收集后系统如何存储管理，传统图书馆集成管理数据库是结构型数据库，大数据包含有许多非结构化、半结构化数据，这些数据如何进行动态性管理并进行实时分析，如何利用技术进行挖掘、分析，又如何有效进行质量控制等。（3）用户利用信息资源问题。在大数据时代，用户面对海量的数据资源，必受海量数据资源困扰，如何才能快速、有效的获取所需信息资源。（4）图书馆如何获取、分析、管理用户信息行为并加以利用。

3 大数据时代图书馆业务管理与服务创新

3.1 构建信息资源采访数据集成平台

高质量信息资源采访是提高图书馆馆藏质量的关键，是图书馆实现其战略目标的重要保障。高质量的信息资源采访受馆藏信息、用户阅读行为信息、出版业图书出版信息、供应商信息等众多因素的制约。在大数据时代，这些信息多数是非结构化、半结构化的

数据，是孤立的、异构而又分散的。图书馆可以通过某种手段和工具，按照一定的逻辑关系把这些数据集成起来，构建一个统一的、透明的、可检索的平台，可实现资源共享。通过这个平台，图书馆采访工作人员可以实时地动态地了解馆藏情况，用户阅读需求、不同层次用户阅读特点及规律，了解图书市场出版情况，掌握核心出版社图书出版动态，了解供应商实际综合情况等。进行图书采选时利用大数据分析技术，建立图书评价工具模型，根据图书评价体系，各种因子的影响权重，精准地挑选出符合各自图书馆的图书。利用信息资源采访数据集成平台采选图书，更具客观性，避免掺杂太多采访工作人员个人主观色彩。这种购书方式有利于提高图书借阅率，有利于提高馆藏质量，有利于提高文献资源购置费的有效利用。

3.2　用户数据管理

图书馆宗旨是最大程度地满足用户的信息需求，因此，对用户信息的收集尤为重要。传统图书馆对用户信息管理主要针对用户到馆利用图书馆办理图书借阅或刊物阅览，管理信息也是最基本的。对于那些通过网络远程访问或咨询的用户，图书馆不知道他们到底是谁，不知道用户离开图书馆后的行动轨迹，无法进行有效管理。在大数据时代，用户信息是海量的，网络社交、消费行为、电子商务、网络阅读喜好等大量粹片式数据存在，这些数据对图书馆研究用户不同人群阅读特点及规律极为重要。图书馆可以充分利用大数据分析平台，把用户数据和信息结构化的、非结构化的整合到统一的平台，构建全面的用户视图，利用大数据分析技术、聚类技术和挖掘技术，通过对用户数据进行分析、聚类和挖掘，从而有效甄别出优质用户、潜在用户和流失用户。此外，在数据时代，用户只要发生过行为，就会留下痕迹，如何保护用户的隐私权，就摆在图书馆人面前。

3.3　大数据存储管理

在大数据时代，数据存储需要面对两个问题，即存和用。所存的数据要完整，要持久，用户使用数据要及时，有效。大数据巨存量对传统图书馆 IT 架构、存储手段提出巨大挑战。传统图书馆保存纸质图书只有两三百年，且还跟环境和保存手段有关，馆藏资源数字化后可以保存持久。目前图书馆正处于向数字化图书馆转型期间，大量纸质图书有待数字化，从“数字化”到“数据化”还需漫长过程，馆藏资源数字化是实现未来图书馆在大数据时代战略位置的基础。面对海量大数据，图书馆考虑存储方式时既要考虑实际情况，又要考虑未来发展趋势，选择适合自己的介质。按照某种标准，如学科专业或地理区域，把大数据划分成一间间区域，借助非关系型的数据库分析技术，利用颁布式存储技术存储，使图书馆海量数据更加完整有序。此外，图书馆在存储大数据时要注意数据的清洗和过滤，去掉“脏”的数据，使存储数据保持高质量。关于大数据存储问题，国家图书馆较早关注。国家图书馆一期维修改造完成后，非结构化数据存储量将达到 800TB 左右。

3.4　大数据处理技术和分析技术

传统图书馆对数据的处理主要是将纸质文献资源数字化、网络化，建成关系型数据库以满足用户的信息需求。在大数据时代，图书馆对数据的处理方式、范围、对象等将发生巨大的变化，传统的数据处理技术无法处理异构性数据，大数据复杂的数据结构对图书馆技术水平提出更高的要求。目前较成熟的大数据处理技术、分析技术有：Hadoop、SAP HANA、Hive、Pig 等。Hadoop 作为大数据存储与计算软件平台，主要由 HDFS、MapReduce 和 Hbase 组成，是以 Java 为基础的开源软件框架。目前广泛运用于谷歌、亚马逊、雅虎、Facebook 等领域。HDFS 是大数据存储平台，MapReduce 是简洁的面向大数据分析和处理并行的计算模型。MapReduce 技术是非关系数据管理及分析技术的典型代表。Hive 是数据仓库平台。通过 Hive 平台，技术人员可以进行海量数据提取、转化、加载（ETL）工作。Pig 是大规模数据分析平台。通过 Pig 平台，程序员可以进行大规模数据处理。Hive 和 Pig 都提供类 SQL 语言解决方案，减少了程序员的工作难度[5]。

3.5　服务创新

图书馆核心价值是为用户提供信息服务。大数据给图书馆带来丰富的数据资源，为图书馆尽最大可能满足用户信息需求提供坚实物质基础。在大数据时代，用户信息需求特点、用户服务方式、模式将发生重大改变。图书馆抓住大数据机遇为用户提供更好地信息服务：一是提供数据资源一站式服务，图书馆要转变传统服务理念，以个性化，智慧化为服务理念，把网络资源、数字化资源、馆藏资源、学术资源等整合起来，构建统一数据资源服务平台，提供无缝隙一站式服务。二是提供增值服务，利用数据挖掘技术，结合优势专业队伍，挖掘大数据“矿藏”，分析、归类数据，发现数据知识，从而提高数据的价值密度，以满足用户信息需求为出发点来提供增值服务；三是提供智能化、智慧化服务，图书馆利用智能代理技术等大数据技术，遵循用户活动轨迹，描述用户活动行为，就可以准确确定用户信息需求，为用户提供智能化、

智慧化服务。图书馆也可以提供“私人定制”服务。

3.6 人才培养

大数据利用效果如何关键在于人。在大数据时代，图书馆工作内容、业务管理模式、用户服务方式等都发生了变化，需要具备大数据相关知识，精通数据技术，发现数据知识、善于数据管理等方面的人才，而这样人才全世界都很稀缺。麦肯锡 2011 年 6 月做报告时称，美国到 2018 年将缺乏具有“深度分析”经验工作者 14~19 万名，精通数据经理人 150 万名。图书馆可在现有基础上，培养既有图书情报专业知识，又具有数据分析、挖掘数据价值专业人才。一是通过人才委托培养方式，委托成熟的专业处理公司培训，提升工作人员处理数据能力。二是通过高校与大企业开展联合大数据教育方式，培养高素质的数据人才。

参考文献：

[1] 但彬.大数据=海量数据+复杂类型的数据[EB/OL.[2012-07-12].http://www.dlnet.com/news.hyxg/88831.htm.

[2] Philip Russom.big data analytics[EB/OL].[2012-08-01].

[3] 徐子沛.大数据：正在到来的数据革命，以及它如何改变政府、商业与我们的生活[M].桂林：广西师范大学出版社,2012：40-57.

[4] IBM 公司在大数据领域占有先机 [OL].[2012-08-01].http://it.hilizi.com/server/275232/372589013274b.shtml.

[5] 李璠,贾鸿飞.大数据时代银行业的机遇与挑战[J].中国金融电脑,2012,(12):25-29.

[6] 朱静薇,李红艳.大数据时代下图书馆的挑战及其应对策略[J].现代情报,2013,(5):9-13.

[7] 黄晓斌,钟辉新.大数据时代企业竞争情报研究的创新与发展[J].图书与情报,2012,(6):9-14.

[8] 韩翠峰.大数据时代图书馆的服务创新与发展[J].图书馆,2013,(1):121-122.