大数据时代的图书馆科研用户服务模式探索

中科院国家科学图书馆成都分馆
周涛 杨志萍 王春明
通信地址：成都市一环路南二段中科院成都文献情报中心，610041
邮箱：zhout@clas.ac.cn

摘要 大数据时代的来临，图书馆特别是研究型图书馆以及大学图书馆正面临着贡献边缘化的危机。通过对知识创造的生命周期模型进行分析，当前和未来科技创新需要科研数据管理和基于知识的交互协同创造能力。图书馆服务应抓住机遇，通过科研数据管理与用户关系管理相结合，探索融入科研一线，跟踪科研全过程的图书馆知识化服务模式，提升图书馆的竞争力。本文基于国外的图书馆科研数据管理以及用户关系管理方面，从技术支撑、科研数据组织、数据分析到用户关系管理等，探索该服务模式。
关键词 大数据，图书馆服务，科研数据管理，用户关系管理

ABSTRACT In the Era of Big Data，libraries, especially research and university libraries, are facing threats to their contributions. Based on analysis of the life cycle model of research knowledge creation, research data management and knowledge-based are identified as challenges to library services. Libraries have to grasp the opportunity to develop e-Science knowledge management by combing research data management and Customer Relationship Management (CRM). Based on the research data management and CRM experience in foreign libraries，this article explores the service in libraries from these aspects: technology support, research data organization, data analysis and CRM in libraries.

KEY WORDS Big data, Library service, Research data management ，CRM

大数据(Big data)是 IT 界继 Web2.0、云计算之后近两年最流行的词。数据目前尚没有统一的定义，通常被认为是一种数据量很大、数据形式多样化的非结构化数据[1]。奥巴马政府于 2012 年 3 宣布推出联邦政府“大数据的研究和发展计划”，旨在推进和改善联邦政府部门的数据收集、组织和分析工具及技术，以提高从大量、复杂的数据集中获取知识能力[2]，把大数据上升到了国家战略的高度。随着大数据时代的来临，科学研究也正在向数据密集型科研转变[3]。对于图书馆特别是研究型或大学图书馆来说，在这个“大数据”时代如何提高海量增长的文献数据处理能力等，尤其是科研用户服务能力，是图书馆研究的思考之一。解决这个问题的关键之一就是图书馆要构建强大的科研数据管理以及科研用户关系管理有机结合的管理体系，实现图书馆，特别是研究型或大学图书馆的科研用户服务能力的提升，推动图书馆服务事业的发展。

1. 图书馆科学数据管理服务

科研是与数据密不可分的，科研的过程就是数据的发现、收集、处理、分析以及利用等的过程（图 1）。科研数据具体是指数字形式的研究数据，包括在研究过程中产生的能存贮在计算机上的任何数据，也包括能转换成数字形式的非数字形式数据[5]。包括科研论文、专利、研究报告、实验观测数据和元数据、参考资料、照片和图表、学术类多媒体资源等等。

近年来，随着各国的科技投入增大，科学观测和分析能力已得到快速的提升，导致科研数据的产生和积累呈指数级增长。有效的科研数据管理具有保护数据免于丢失、提高数据曝光度，传播和出版成果、实现数据共享、对科学质疑公开、鼓励观点的多样性、节约科研成本、完成研究资助方的要求等诸多意义。

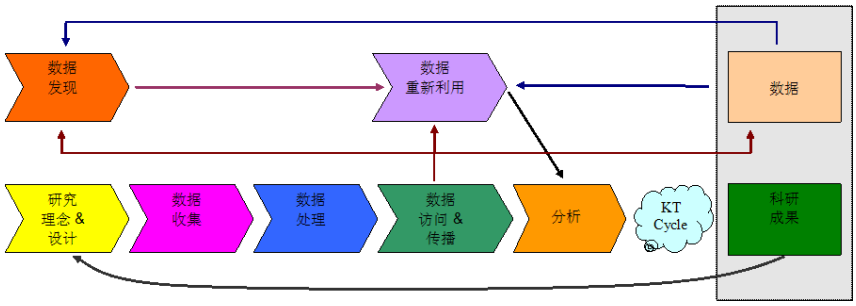

图 1 科研知识创造的生命周期模型[4]

目前在美国科学数据管理成为美国研究型大学图书馆的一项新使命，美国多家知名图书馆都积极投入。2011 年美国国家科学基金会（NSF）要求所有基金申请必须提交研究数据管理计划，包括数据的长期保存、共享与访问方式等内容。这项战略性信息基础建设新政策强调了公共获取数据的重要性，并且 NSF 的 datanet 项目明确研究型图书馆将作为主体参与此项工作。科学数据管理离不开信息技术的支撑，需要依托一定的基础设施和软件平台。

1.1 基础设施

1.1.1 图书馆自建或与其他机构合作建设

图书馆独立建设数据仓储并对其进行维护与管理，目前这种形式并不多，这并不仅仅是图书馆的技术能力问题，在开放、共享的理念下，图书馆更多地趋向于合作方式，重点利用学校建立的数据或机构仓储，利用已有的基础设施，与校园内其他部门或政府和一些组织资助的项目也建立了不同学科的数据仓储，为国内机构提供共同服务，使科研数据存储超越了图书馆，超越了某一个单一机构。图书馆这时的任务是向研究者提供相关信息与帮助，使其了解这些仓储，并帮助用户利用这些重要资源。在技术上侧重科研环境建设，构建数据门户，做好数据导航。

1.1.2 利用云计算共享

近几年来，“云计算”的应用在图书馆领域也迅速发展了起来。云计算会促进 e-Science 发展，通过云远程使用资源，图书馆不需要自己购买硬件设施，不需要建设基础设施环境，大型设备尽可在网上利用，图书馆可以将数据存储在云存储服务商提供的服务器中，按需申请，按时付费。例如微软的云计算战略以及云计算平台——Windows Azure 以云技术为核心，提供了软件+服务的计算方法。 它是 Azure 服务平台的基础。Azure 能够将处于云端的开发者个人能力，同微软全球数据中心网络托管的服务，比如存储、计算和网络基础设施服务，紧密结合起来。这样开发者就可以在“云端”和“客户端”同时部署应用，使得企业与用户都能共享资源。

目前，云计算在图书馆的应用越来越广泛。OCLC 启动的“将图书馆管理服务推向 Web 级的战略”是图书馆界接受云服务的重要标志性事件[6]，美国国会图书馆与 DuraSpace 正式发布的开源云服务 DuraCloud 项目部分受美国国会图书馆的全国数字信息基础设施与保存项目（NDIIPP）资助[7]。DuraSpace 项目开源，但提供基于订购的服务。目前已有麻省理工学院、哥伦比亚大学、西北大学和莱斯大学签约使用其托管的云服务以保护数字资源。

1.2 科学数据组织服务

研究人员的科研数据除保存在相关学科库以外，机构仓储是另一重要选择。机构知识库

最初的设想是保存机构成员的研究成果，并提供出版机会，既有存储的功能，又有检索和服务的功能。机构库的创建软件多种多样，目前国际上流行的软件平台是Eprints以及Dspace[8]。

1.1.1 Eprints

机构知识库系统的发展始于2000年英国的南安普敦大学开发的Eprints软件，南安普顿大学采用自己开发的Eprints软件创建的遵循OAI协议的机构知识仓储，该机构知识仓储中目前保存的数据的类型收录的学术内容格式既包括结构化的。

1.1.2 Dspace

美国麻省理工学院(MIT)图书馆和美国惠普公司实验室合作两年多于 2002 年发布全球第一个机构知识库(IR) —DSpace 数字资源存储系统，并将其 BSD 开放源代码技术向全球公开。DSpace 在目前的数字仓储软件中占据了三分之一以上的份额。作为开放源代码，它允许被下载、修改，而且其所使用的第三方软件也都是开放源代码系统。DSpace 可保存任何格式的数字资源，包括论文、图书、图书章节、数据集、学习资源、3D 图像、地图、乐谱、设计图、预印本、录音记录、音乐录音、软件、技术报告、论著、视频、工作文档等。MIT 图书馆数据管理项目组承担全部数据的存档、管理、系统维护、软件升级和用户使用指导等服务并且与出版商争取相关权益，建立开放获取政策，执行 DSpace 数据提交服务，推动 MIT 的开放获取服务等。另外在 DSpace 的众多用户中，剑桥大学的机构仓储较为成功。

然而调查显示，国际上流行的软件平台 Dspace、Eprints 软件在国内并不十分受欢迎。有的原来使用 Dspace 的机构库也在运行过程中慢慢更换了软件平台比如中国科学院国家科学图书馆，中国西部环境与生态科学知识积累平台。

1.2 科研数据分析服务

科学数据的组织管理服务与图书馆的其他资源服务想类似，是图书馆开展科学数据服务基础。未来图书馆科学数据服务大趋势是服务中附加更多智力活动，进行数据分析，把科学数据进行关联，帮助用户更好地利用数据。

目前逐渐有一些机构仓储在存储数据的同时提供了类似的服务。康奈尔大学组建了研究数据服务组（research data management service group），其图书馆作为其中主要成员与校内其他机构合作，提供各种研究数据管理服务，包括存储备份、元数据加工、数据分析、数据发布、协作工具等[9]。康奈尔大学图书馆近两年正在探索开发研究数据检索挖掘工具，建立一套标准符合 NSF 要求的数据管理与服务体系，并且已经建立了一个实验性的数据仓储 datastar[10]。Datastar 目前主要保存农业与生态系统学科的研究数据，支持研究合作与数据共享，促进研究数据及其高质量元数据的发布存档。哈佛大学的“dataverse 网络”(Dataverse Network) ，项目包括科研数据出版、共享、参考、抽取和分析各个方面，为大学或其他机构提供数据出版系统的全部解决方案，并提供数据分析服务。目前可提供数据分析的机构不多，这是图书馆科研数据服务的方向。

2. 图书馆用户关系数据管理

用户是数字图书馆建设的出发点，为了提升数字图书馆的服务质量，需重视数字图书馆的用户研究。用户关系管理（Customer Relationship Management，CRM）是通过有关的管理技术和方法对用户进行系统化研究，识别有价值的用户，对用户进行沟通和教育培训等工作，从而改进服务，提高用户满意度。

数字图书馆用户关系管理借助数字仓库、数据挖掘、知识发现、专家系统和人工智能等

多种现代信息技术手段，建立一个能搜集、追踪和分析用户信息的系统，为数字图书馆用户服务和决策提供一个自动化的解决方案，实现数字图书馆由传统的人工管理模式向现代管理模式的转变。实际上，目前有些数字图书馆系统本身就具有用户数据的自动收集、统计和分析功能。数字图书馆用户关系管理的一些新技术，如数据仓库技术、数据挖掘技术和知识发现技术等，有效地使数字图书馆用户数据的获取、模式发现、数据的积累、传播和共享更为快捷有效。

要做好用户关系管理，就需要搜集各种用户数据信息，对用户资料进行统一管理，包括用户基本信息、用户类型划分、用户状态、服务情况等信息进行整合。数字图书馆数据挖掘的信息源主要是用户活动信息、日志文件、网站的注册用户信息。通过整理和分析日志文件可以获得许多有意义的信息。如页面访问量、受欢迎程度等，或者了解到用户的爱好、价值取向，从而为制定有关的服务策略提供依据。用户使用数据的挖掘分析有许多方法，如关联分析是为了挖掘隐藏在数据间的相互关系，序列分析的侧重点在与分析数据的前因后果或顺序关系等。用户的数据挖掘不仅可以了解数字图书馆的访问量，而且可以统计不同的数字化资源被访问的频率等，还可以对用户进行跟踪分析，从不同的侧面来研究用户的信息需求及其行为规律。

用户关系管理(CRM)系统应该具有用户信息分析能力、集成能力和用户互动渠道。通过建立数据仓库对大量用户数据进行综合分析，识别相关用户群，根据多种指标对用户进行分类，针对不同的用户实施不同的策略，为用户提供更合适的服务。

目前国外有许多图书馆引进ＣＲＭ系统。大英图书馆是一个世界上最大的研究图书馆，其读者遍布世界各地。维护其客户和图书馆积累的大量数据，是一项重大挑战，该组织发现自己与 37 个不同的客户相关的数据库。为了把一个系统在其 37 数据库中的信息集成，大英图书馆引进安装 Microsoft Dynamics CRM（MSDN）。另外芝加哥图书馆用户关系管理采安装了具有类似功能的 SageCTM 系统。 SageCRM 客户服务系统可以提供完整的工作流、问题跟踪、案例管理及服务状态的信息，帮助创建一个可靠的知识库，从而保证一致而高效率的客户服务。

3. 基于科研数据管理与科研用户关系管理的服务模式思考

图书馆特别是研究型图书馆的发展在从传统图书馆转变为数字图书馆之后，未来的发展之路是要发展嵌入科研一线的知识化服务模式[11]。嵌入科研一线的知识化服务强调了图书馆必须更为直接地服务于科研人员。 在数字科研下，研究数据管理保存数据集存储与分析对支持重复验证、全面传播知识、激发新问题、开展研究等都有重要的作用[12]。而图书馆为了更好地实现以需求为驱动的服务模式，建立图书馆自己的用户管理系统也十分的必要。

研究型图书馆为了更好地服务科研人员，把科研数据管理与用户关系管理结合起来显得十分必要。为此可以建立一个具有科研用户与科研数据之间、科研数据与科研数据之间、科研用户数据与科研用户数据的数据系统（如图 2）十分必要。采用关联分析以及序列分析等，从不用侧面挖掘与跟踪某一个或几个科研用户，或者某个学科领域科研用户在科研过程及之后的用户需求。

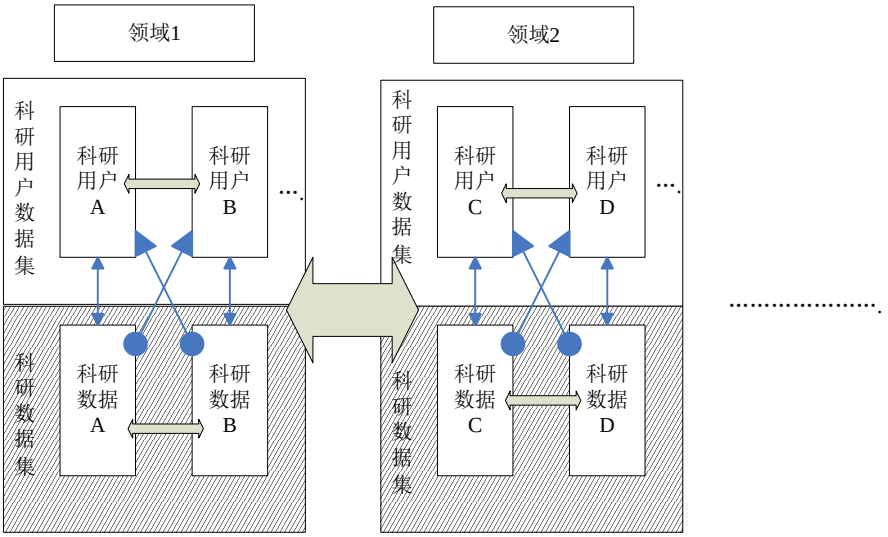

图 2[1] 科研用户数据[2]与科研数据系统

参考文献：

[1] 大数据时代的特点[EB/OL][2012-05-20]
http://www.51lian.cn/html/2012/xueshu_0417/32237.html

[2] 赛迪智库软件与信息服务研究所，美国将发展大数据提升到战略层面[J],中国电子报，2012-07-17(003).

[3] Hey T, TamSley S, Tolle K. The fourth paradigm-Data-intensive scientific discovery[OL]. [2010-09-14].
http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[4] Humphrey, Charles. (2006),e-Science and the life cycle of research, http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

[5] ANU Data Management Manual: Managing Digital Research Data at the Australian National University [2010-09-1], http://www.citeulike.org/user/janeta/article/10426141

[6]OCLC 宣布将图书馆管理服务移到“网络规模”的策略
[OL].[2010-09-14].http://www.oclc.org/asiapacific/zhtw/news/releases/200927.htm

[7] DuraSpace 推出开源云服务 DuraCloud
http://www.duraspace.org/duraspace/launches/open/source/cloud/service

[8]纪云霞，国内机构库软件平台调研，图书情报工作网刊，2012（1）

[9] Research Data Management Service Group(RDMSG)[EB].[2012-3-15]. https://confluence.cornell.edu/display/rdmsgweb/home;jsessionid=73DF1608333FB2D6FoFDCB976AB20C76.

[10]康奈尔大学实验性的数据仓储 datastar，
http://datastar.mannlib.cornell.edu/

[11]张晓林，研究图书馆 2020：嵌入式协作化知识实验室，中国图书馆学报，2012（1）

[12] Christine L Borgman. The conundrum of sharing research data[OL].[2011-11-15].

[1] 本图的科研数据 A 是有科研用户 A 的科研活动产生。以此类推。
[2] 此处的科研用户数据是指科研用户的基本信息、类型划分、状态、服务情况等等。

http://ssrn.com/abstract=1869155

作者姓名：周涛 杨志萍 王春明
单位：中科院国家科学图书馆成都分馆
通信地址：成都市一环路南二段中科院成都文献情报中心，610041
联系电话：028-85223722
邮箱：zhout@clas.ac.cn