加州大学数字图书馆的数据驱动战略实践与启示*

王茜[1,2] 孙蒙鸽[1,2] 郑新曼[1,2] 刘细文[1,2]
（1. 中国科学院文献情报中心，北京 100190；2. 中国科学院大学信息资源管理系，北京 100190）

摘要：当前国内学术图书馆正处于积极向数据驱动服务转变的时期。以美国加州大学数字图书馆建设为例，梳理其数据驱动战略的实施框架和具体实践，对其数据驱动工具开发、数据基础建设、数据驱动服务进行详细的调研与分析，在此基础上获得三点启示：加强数据管理流程的构建、数据驱动工具的开发；嵌入科研一线并进一步推动学术交流；关注多形态数据建设、建立多元合作渠道。期望为我国学术图书馆的数据驱动服务转型提供参考。

关键词：数据驱动战略；数字图书馆；图书馆服务；加州大学；学术图书馆
中图分类号：G250.76；G250.7；G250.71　DOI：10.3772/j.issn.1673-2286.2023.11.002
引文格式：王茜，孙蒙鸽，郑新曼，等. 加州大学数字图书馆的数据驱动战略实践与启示[J]. 数字图书馆论坛，2023（11）：10-19.

随着大数据技术的成熟，数据驱动深刻影响着每一个行业的发展，已然成为第四种科研范式[1]。数据驱动模式是现代组织机构面对大数据时代海量数据而提出的转型思想，将成为获得竞争优势和快速发展的关键基础[2-4]。在数据密集型科学发现的激发下，数据成为人们认识、分析与解决问题的中心，这使人们开始利用数据思维解决现有问题[1, 5-6]。数据驱动指利用技术手段采集获取数据，并将数据加工、组织，转化为信息，随后对信息进行分析、整合和提炼，形成知识，最终训练生成自动决策系统并提供服务的过程[5, 7]。数据驱动的基础是数据，数据驱动的核心是发挥数据动能[5]。虽然目前对于数据驱动的定义还没有一个统一的说法，但总的来说，数据驱动是以数据为基础，以搭建数字化业务流程为目标，以数据计算为支撑，最终形成数据到业务目标的反馈循环的工作模式。数据驱动包括了从数据到信息，从信息到知识，以及从知识到决策的转换过程。数据驱动使对海量、多元且异构数据的获取、分析、挖掘、保存、管理等成为了现实。

在数据思维模式下，将数据驱动应用于不同场景，能够输出数据驱动的设计、数据驱动的程序、数据驱动的产品等成果[5]。数据驱动的图书馆服务就是将数据驱动应用于图书馆建设，以提供更好的服务，包含数据汇聚和管理（服务基座）、数据驱动工具（服务能力）、数据驱动服务（智慧服务）3个部分。数据汇聚和管理是图书馆的基础底座，数据驱动工具是数字图书馆服务能力的体现，它们共同帮助实现了数字图书馆的数据驱动服务。

本文以美国加州大学数字图书馆（California Digital Library，CDL）的数据驱动战略为例，将数据驱动型图书馆服务的构建过程与数据管理流程相结合，对CDL的数据收集、数据驱动工具研发、数据驱动服务等进行系统调研，期望其能够作为一个典型案例为国内学术图书馆的数据驱动服务转型提供参考。

收稿日期：2023-09-27
*本研究得到国家自然科学基金重点项目“数智赋能的科技信息资源与知识管理理论变革”（编号：72234005）资助。

1 CDL的数据驱动战略框架

CDL由加州大学于1997年创立，旨在利用新兴技术改变数字信息管理和获取方式[8]。CDL以校园内部合作为基础，外部合作为扩展，扩大了其学术与资源的影响力，最终提供变革性的数据驱动型图书馆服务，为出版、共享和保存学者日益多样化的学术成果，提供丰富、直观和无缝的数字化环境，改变了教师、学生、科研人员等用户发现、获取信息的方式。CDL遵循开放性、多样性、以用户为中心、合作性、适应性、敏捷性以及创新性原则，构建了一个不限信息访问，支撑技术创新的数字图书馆。CDL利用自身所包含的资源、技术、合作伙伴多等优势，迅速创建大量、多模态的馆藏数据基底；加快构建集成化、便捷化的数据驱动工具；积极开展智能化、个性化、创新化服务，形成了以多样化元数据为基础、以智能化技术为支撑、以广泛的合作为辅助、以促进学术交流为目的、以满足用户需求为发展要求的数据驱动图书馆服务模式。

CDL构建了集成化的数据驱动工具，保存了大量的数字资源，并开展了数据驱动的服务（见图1）。数据驱动工具是CDL的核心支撑部分。根据服务与使用对象，将CDL数据驱动工具分为帮助构建服务基座（数据基础）的面向管理者的工具，以及帮助建设数据驱动服务的面向用户的工具两大类。通过这些工具，数据在CDL中转换与重组。首先CDL管理者收集、检查、处理、描述、管理、集成数据；然后用户在CDL中发现数据，并对发现的数据（或其他数据）进行分析，在分析后将自己的数据提交到CDL进行描述、管理、存储、共享与发布；最后用户发布的新数据（例如对数据进行统计分析的文章）可能又会被收集到CDL中作为馆藏的一部分。在这个过程中，数据实现了在图书馆与用户、用户与用户之间的转换与循环，本文将这个过程命名为CDL数据管理流程。

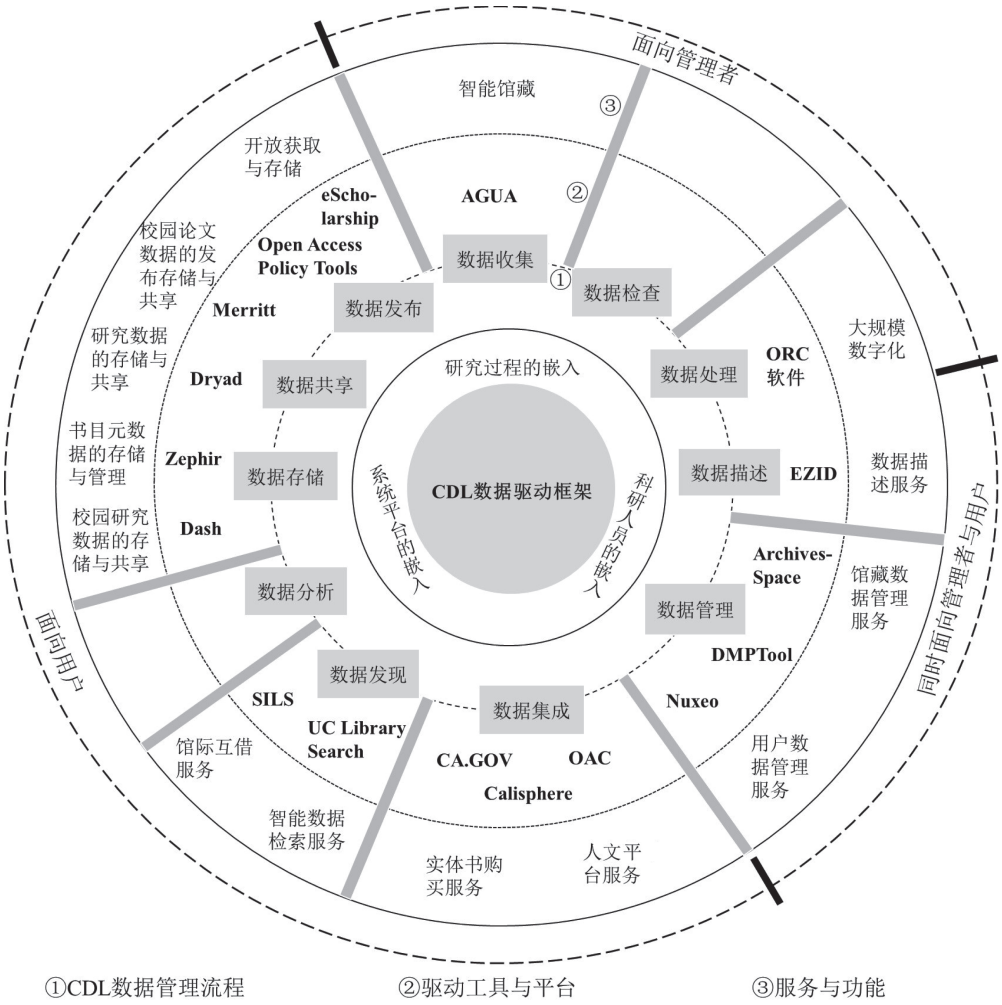

图1 CDL的数据驱动战略框架

根据CDL数据管理流程，对CDL的数据基础、模型工具以及相应服务进行了划分。数据收集、检查、处理属于数据基础设施建设中必不可少的环节，CDL通过创建面向图书馆管理者的数据收集、数据处理等工具，提供了智能馆藏、大规模数字化等服务。数据集成、数据发布、数据存储、数据共享、数据发现则可被视为CDL专门为用户提供的服务，CDL通过Dryad、Dash、eScholarship等平台与工具为用户提供了数据驱动的人文平台、研究数据存储与共享、开放获取与存储等服务。此外，在CDL数据管理流程中，数据描述与数据管理则可被视为同时面向管理者与用户的服务，它既能帮助CDL管理馆藏数据也为用户描述、管理个人数据提供便利。

虽然CDL提供了一些分析工具（如File-Analyzer）[9-10]，在开展馆藏数据收集等时也会进行数据检查，但目前CDL平台中还缺少数据检查与数据分析方面的开放型集成智能工具、平台与服务。就整个CDL数据管理流程而言，CDL提供了较为全面的数据驱动工具平台、数据基础、数据驱动的服务与功能，通过这种方式将系统平台、研究过程以及科研人员关联在一起，促进了学术的交流与数据的再利用，推动了学术信息的数字化转型[11]。

2 CDL数据驱动工具的研发

CDL的数据驱动工具为数据在CDL与用户间的循环流动提供了动力，并为数据驱动服务的构建提供了支撑。根据面向对象可将CDL研发的数据驱动工具分为面向管理者、面向用户以及同时面向管理者与用户的数据驱动工具三大类，其中面向管理者的模型工具主要用于数据基础建设，面向用户的模型工具则被用于构建数据驱动服务，同时面向两者的模型工具则具备两种功能。对CDL的数据驱动工具进行调研与分析能够帮助理解CDL的数据驱动服务战略思路。

2.1 面向管理者的数据驱动工具研发

2.1.1 数据收集工具研发

馆藏资源建设是图书馆的基础工作之一，满足用户需求是图书馆馆藏优化的目标[12]。为了让馆藏决策变得更智能，馆藏内容更丰富，CDL开发了一种基于Web的数据驱动的智能馆藏决策技术AGUA。AGUA支持美国西部学术资源区域合作储存项目（Western Regional Storage Trust，WEST），它能够收集成员机构以及外部的数据，对上述数据进行识别，根据标准组合计算风险，确定归档的最佳候选者，并规定归档期间应采取哪些行动，以加强归档的可靠性和准确性。WEST能够大规模地执行此分析工作，从数十个成员和外部机构中获取资源，并向拥有最多资料、最佳存储条件的成员机构提出建议。WEST还能够促进成员之间的合作，通过比较成员的馆藏并在AGUA分析生成的报告中强调成员之间的联系来填补存档馆藏的空白。AGUA的馆藏比较、馆藏分析等功能为CDL及相关合作成员进行高效馆藏数据收集与存储提供了便利，为进一步开发数据驱动工具、实现数据驱动的服务提供了基础。

2.1.2 数据处理工具研发

CDL可通过数字化技术将大量的实体馆藏材料转化为数字化格式，使用户通过数字平台即可访问和利用这些内容。CDL实施大规模数字化服务，通过逐页拍摄书籍，使用光学字符识别（OCR）软件，最终生成可搜索的文本。截至2023年8月，大规模数字化内容共包括1 446 801个全视图卷、4 670 531册书籍[13]。数字化内容可存入HathiTrust数字图书馆、Google图书馆项目和当地校园数字化计划，以方便内容被更广泛地访问和利用。CDL大规模数字化服务的目标包括辅助学生和教师开展科研、开辟学术探究的新领域、履行公共服务使命、保护和高效管理馆藏内容，以及提高馆藏管理效率等[13]。大规模数字化服务改变了文献、书籍等数据的存储方式，间接改变了用户的访问模式，直接推动了访问工具（如检索平台）的进一步创新；为CDL解决了很多问题，包括提高了可及性、方便用户获取知识、促进学术研究以及提高图书馆效率等；为加州大学的学生、教师和研究人员提供了一个全新的数字化信息空间，为学术探究提供了更多支持；使公众能够更轻松地访问加州大学图书馆资源，履行图书馆的公共服务使命；帮助确保加州大学图书馆馆藏的长期可用性，并提高馆藏管理的效率。总之，实体馆藏的大规模数字化，是CDL数据驱动工具平台以及数据驱动服务的基础之一，为后续工具的研发与服务的制定提供了可能。

2.2 面向用户的数据驱动工具研发

2.2.1 书目元数据的存储工具研发

在书目元数据的存储与管理方面，为了创建一个能够存放大规模数字化书籍数据的学术图书馆馆藏环境，CDL提供了一个名为Zephir的系统，该系统能够从HathiTrust平台[14]获取、存储和管理书目元数据，并导出这些数据以用于其他HathiTrust系统。HathiTrust是机构合作委员会（Committee on Institutional Cooperation，CIC）与加州大学图书馆在2008年创建的数字图书馆项目，旨在将其成员所收藏的纸质文献进行数字化存储，为用户提供数字服务。该平台中保存了超18 000 000份数字化资料，并提供美国和国际版权法允许的最大限度的阅读访问权限。用户可以将数字化材料存放在HathiTrust中，以便长期保存和访问。

2.2.2 科学数据集的存储共享与发布工具研发

科学数据集具有不可估量的价值，如果没有适当的文档记录、长期存储以及可发现与可访问路径，它们就会随着时间的推移而失去价值。大量小型异构数据集通常被存储于表格中，这种存储方式会使许多数据集面临失去价值或丢失的风险。2011年CDL与微软研究院（Microsoft Research，MSR）以及戈登和贝蒂摩尔基金会（Gordon and Betty Moore Foundation）合作，为Microsoft Excel软件创建了DataUp数据管理工具，并为用户提供了数据驱动的研究数据的发布、存储与共享服务。本着让研究人员能够更加轻松地描述、存储和公开分享他们的研究数据的目的，加州大学创建了一个名为Dash的开源平台。由于Dash和DataUp的功能逐渐重叠，2014年CDL对两个平台进行了合并，并最终仅保留Dash平台。Dash平台能够科学组织任何领域中的任何类型数据、检索与重用数据、将数据连接到相应的文章以及资助组织，有效地促进了研究数据的共享和获取。但需要注意的是，Dsah平台的服务对象主要为加州大学校园内部的用户。

除了专门为校内用户提供科学数据集发布、存储与共享服务外，CDL还提供了一个开放的数据发布平台Dryad。2018年，CDL与科学数据库机构Dryad合作[15]，实施了“Dryad-CDL计划”，为CDL平台增加了面向公众用户的数据驱动的科学数据发布、存储与共享服务。Dryad最早于2007年，由国家进化综合中心（NESCent；当时是杜克大学、北卡罗来纳大学教堂山分校和北卡罗来纳州立大学的合作项目）在美国国家科学基金会的资助下创建。后于2011年提出了联合数据归档政策（Joint Data Archiving Policy，JDAP），并提出科学数据也应被存储于公共设施中这一理念。Dryad与学术期刊积极合作，研究人员在学术期刊上发表的数据可同步提交到Dryad中，从而实现了学术期刊与科学数据库的连接。根据Dryad介绍，截至2022年，Dryad平台与70 000余个国际机构、1 000多种学术期刊进行了合作，集成了50 000余种数据出版物[16]。

2.2.3 学术出版数据的存储共享与发布工具研发

在数字化背景下，传统的学术出版模式已经难以适应新的需求，而目前存在的电子出版平台又具有付费高、出版体系不合理等问题，因此图书馆与出版商均需要寻找能够适应数字化时代需求与趋势的新学术出版平台。CDL建立了一个开放学术出版平台eScholarship[17]，提供了一系列开放获取资源、学术出版服务和研究工具，支持开放获取原始数字出版期刊、图书、论文、会议论文集等，科研人员、科研单位和出版商可以在该平台上进行成果发表、数据检索、内容浏览、评论和保存数字化成果等操作[18]。eScholarship支持的学术出版模式包括：预印本、数字出版成果、既面向研究人员又面向大众读者的电子学术著作等[19]。eScholarship除了是一个开放学术出版平台外，还是CDL数据存储服务的重要组成部分。存储的数据包括预印本、工作论文、电子论文、学位论文、学生重大项目和研讨会/会议论文集。自成立以来，eScholarship存储了超300 000种出版物，浏览量超1亿次，内设92种期刊，并存储了超53 715篇论文（包括加州大学10个校区和附属研究中心的工作论文、电子论文和学位论文）[20-21]。

2.2.4 校园数据的存储共享与发布工具研发

除了不同数据的存储、共享工具外，CDL还提供了一个开源新型数据库服务系统Merritt。该系统能够保存和访问加州大学10个校区的各类数据（例如图片、视频、数据集、文本等），供校内图书馆、档案馆、博物馆、学术部门、实验室和其他组织单位使用。基于微服务、关联数据和REST框架[18]，Merritt为用户提供了数

据标识符创建、人工审核、REST应用程序接口、元数据目录制定和数据使用协议制定等功能。目前Merritt拥有400多个馆藏，其中包括加州大学图书馆特别馆藏内容、eScholarship期刊出版物以及来自加州图书馆等机构的数字内容。Merritt还充当了eScholarship和Dryad的保存和访问存储库，2018年CDL宣布Merritt与Dash平台均获得了CoreTrustSeal的认证[22]，2022年Merritt还获得了莫伊根·阿米尼卓越运营奖[23-24]。目前Merritt仍在不断进步，其使用成本从650美元/TB降低到150美元/TB，存储量已从2019年单个对象副本的120 TB增加到2022年的295 TB[25]。

2.3 面向管理者与用户的数据驱动工具研发

2.3.1 数据标识符创建工具研发

开放数据、开放科学、开放获取倡议迅速兴起，图书馆的角色正在快速转变，学术交流（包括数据交流）不断发展，数字存储成为各国关注的话题，因此如何更好地进行数据描述成为一大难题。基于此，CDL提供了EZID工具[26]以实现数据描述，使加州大学的学者和研究人员能够轻松地创建和长期管理具有全球唯一性的标识符数据，并确保数据能被长期发现。目前EZID支持两种标识符：Digital Object Identifiers（DOIs）和低成本的Archival Resource Keys（ARKs）[25]。通过对两种标识符建立联系，用户能够追踪目标数据[27]。目前已有包括加州大学10个分校区以及其他学校或机构（例如密歇根州立大学图书馆、西北大学等）在内的30多个机构应用此工具。除了能够帮助科研人员更好地公开数据、帮助资助者节省数据管理与跟踪的时间外，EZID还能帮助CDL自身更好地管理、存储数据，并成为数据引用、共享与管理的首选场所。

2.3.2 数据管理工具研发

目前，越来越多的研究人员需要参与数据管理活动，基于此CDL提供了一款主要服务于美国研究人员、组织者、资助者用户，能够帮助创建数字管理计划的数据管理工具DMPTool。作为在线数据管理计划工具，DMPTool主要包含两个功能模块：Funder Requirements模块和Public DMPs模块，分别负责汇集数据管理政策与要求、共享数据管理计划。其中：Funder Requirements模块主要汇集了美国各政府部门与资助机构对数据管理的相关政策与要求，方便科研项目工作者获取这些信息，从而成功获得基金的支持；Public DMPs模块则汇集了由公众创建，并得到公开传播许可的数据管理计划[28]。2015年，DMPTool被大数据、伦理与社会委员会（Council for Big data，Ethics，and Society）评为最杰出的高校研究型图书馆所提供的数据管理工具[29]。截至2023年8月，DMPTool使用用户量达102 171位，参与机构有386家，共包含98 511份数据管理计划[30]。

对于数字图书馆内部的所有馆藏数据，CDL提供了两个管理系统——Nuxeo和ArchivesSpace，以管理独特的主要馆藏资源。Nuxeo是CDL的数字资产管理系统，供CDL和指定的附属机构创建和管理对象级元数据和内容文件（如图像、文本、音频和视频）。Nuxeo还支持将数字对象发布到数字文物库平台Calisphere用以访问，以及将数字对象存入Merritt平台用以保存。ArchivesSpace是CDL的档案馆藏管理系统，供加州大学图书馆和加州大学分校图书馆、档案馆和博物馆使用，以维护有关档案馆藏等信息。这两个独立的馆藏系统具有不同的功能和用途，可以管理不同类型的内容，它们共同满足了CDL的不同需求。此外，这两个系统的功能可以相互补充，为用户提供更全面的服务和支持。

3 CDL的数据基础建设

数据是数字图书馆服务的基础，充分获取用户需要的文献、书籍、音频等相关数据在CDL的数据基础建设中占据主导地位。

3.1 数据的来源与类型

CDL的数据来源广泛，可以划分为政府机构、高校及科研机构、用户。其中：政府机构包括加州全州其他图书馆、档案馆、史学会、博物馆和档案馆等机构；高校及科研机构包含WEST成员以及其他合作高校与机构；用户则包括学生、教师、科研工作者等。目前，CDL馆藏资源包含超过1 446 801个全视图卷、4 670 531册书籍、433 115种开放获取学术出版物、92种开放获取期刊、53 715篇论文（包含加州大学内部的工作、学位论文），以及数百万的历史图像、文本、音频和视频，大量站点信息、文物数据等[19, 31]。

3.2 数据的集成

CDL提供了多元而又丰富的数据资源，其中一些具有特色、人文化的数据被集成为多个基于Web的数据驱动的特色平台，包括加州在线档案馆（Online Archive of California，OAC）[32]、网络存档平台CA.GOV[33]以及数字文物库平台Calisphere[34]。这些平台分别以不同的数据内容为基础，以平台为工具，为用户提供了数字驱动的人文平台服务。

3.2.1 档案数据的集成

数字化档案馆平台OAC[32]具有悠久的历史与丰富的内容，以图片或文档的形式收录了来自加州大学的10个分校区以及加州档案馆、图书馆、历史学会、博物馆等多个机构的250 000份特殊藏品。OAC集成了多个机构的数字化档案馆和特殊藏品，具有多语言、高品质的数字化资源和免费的公共访问等服务特色，是一个重要的历史文化资源库。

3.2.2 国家机构数据的集成

网络存档平台CA.GOV[33]存储了数百个加州机构网站。这些网站记录的信息包括国家机构发布的新闻稿、议程、会议记录、活动、报告和统计数据等内容。上述材料通常会随着时间推移而变得不稳定，特别是当领导层发生变化时，一些信息将被隐藏或删除。为了确保这些网站上的信息被保留下来，来自加州大学多个校区、斯坦福大学图书馆等机构的政府信息专家和网络管理员共同创建了CA.GOV，目前该站点共包含1 344条站点信息。

3.2.3 文物数据的集成

数字文物库平台Calisphere[34]收集了来自加州大学10个校区以及全州其他图书馆、博物馆和档案馆等机构的文物资料，包括照片、文件、信件、艺术品、日记、电影、广告、音乐录音等，涵盖了广泛的历史和文化领域，目前其数据总量已超过200万。Calisphere建设的目的是提供免费的、全球范围内的文物数字化资源，以便于研究、教学和一般探索等用途。用户可以通过Calisphere网站搜索和浏览这些文物，也可以下载高清图像和元数据进行研究。

4 CDL的数据驱动服务

通过与CDL数据管理流程相结合，对加州大学为用户提供的数据驱动服务进行划分，可以发现，其为用户提供的数据驱动的服务主要围绕CDL数据管理流程中的数据描述、数据管理、数据集成、数据发现、数据存储、数据共享、数据发布开展。CDL面向用户提供的数据驱动服务（见图2）涉及传统服务、资源管理服务、人文与个性化服务、发布与共享服务。CDL的综合服务较为成熟和完善，对CDL进行调研与分析，有助于为构建我国数据驱动的高校图书馆服务体系提供科学依据。

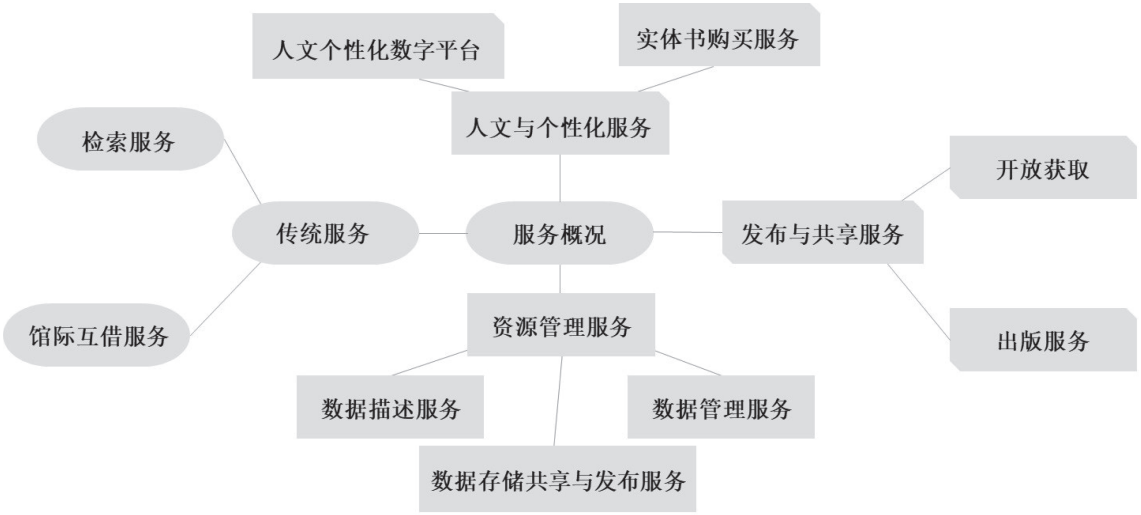

图2 CDL面向用户提供的数据驱动服务

4.1 传统服务

4.1.1 数据驱动的检索服务

以馆藏数据为基础，CDL为用户提供了数据驱动的检索服务。该服务提供了加州大学10个分校检索界面的超链接，每个超链接中的全部资料都可获取，但不同分校图书馆的战略规划文本内容存在差异[35]，例如河滨、洛杉矶等分校均提供出版传播服务，而圣芭芭拉、圣地亚哥等分校不提供出版传播服务。该服务使得用户能够用更短的时间，通过同一个系统在更大的数据范围内查找需要的数据。

4.1.2 数据驱动的馆际互借服务

CDL搜索平台可以链接校外资料，是CDL全系统集成图书馆系统（SILS）[36]的一部分。SILS包含了加州大学10个校区、两个区域图书馆和CDL的数据，做到了跨机构数据的集成，简化了数据搜索的流程，使馆藏管理协作成为可能。通过SILS平台，CDL为用户提供了数据驱动的馆际互借服务，这为用户提供了更轻松的使用体验。

4.2 资源管理服务

4.2.1 数据驱动的数据存储共享与发布服务

图书馆在提供服务时通常单向为读者提供文献、书籍等实体或数据资源，而CDL除了为用户提供自身平台以及其他平台的资源外，还支持用户发布存储自己的数据，下载、共享他人数据，实现了资源在图书馆与用户以及用户与用户之间的流通，并进一步促使CDL平台收集更加丰富、新颖的数据。其中不同的数字资源依托不同的项目平台，这些项目平台包括以书目元数据为基础的Zephir，以科学数据集为基础的Dash、Dryad，以论文数据为基础的eScholarship，以及数据服务系统Merritt。这些项目平台共同为用户提供了多元数据的存储、共享与发布服务，推动并加速了科学交流，为科学研究提供了便利。

4.2.2 数据驱动的数据描述服务

EZID工具通过为数据提供唯一标识，为用户提供了数据驱动的数据描述服务。它对数据集进行了更好的管理与分配，并帮助研究人员分享和获取数据集权限，使数据资源更容易被访问、再利用与验证，这有助于用户规避重复工作，提高科研效率[28]。除了能够帮助对数据进行标注的研究者外，EZID还能帮助资助者根据需求管理与跟踪数据，节省其时间与资源。

4.2.3 数据驱动的数据管理服务

为了方便科研人员进行数据管理活动，CDL开发了DMPTool工具，并为美国研究人员、组织者、资助者提供了数据驱动的数据管理服务。DMPTool具有专业性、共享性、集成性3个特征[28]。专业性体现在DMPTool汇聚了大量政府部门与资助机构在数据管理方面的政策与要求；共享性体现在DMPTool允许有权限的数据管理计划公开传播；集成性则体现在DMPTool支持案例的集成、工具的集成等。DMPTool通过汇集数据管理政策、提供培训平台与共享渠道，为用户申请基金支持、参考他人数据管理计划提供了便利，促进了用户之间以及用户与图书馆之间的信息交流。

4.3 发布与共享服务

4.3.1 数据驱动的开放获取服务

由CDL和校园合作伙伴开发的一套开放获取的学术出版和存储服务工具Open Access Policy Tools，为加州大学开放获取政策作出了贡献[37]。它与多种开放获取期刊合作，为用户提供了数据驱动的开放获取服务，使加州大学部门、研究单位、出版计划和个人学者能够分享并获取开源学术资源。

4.3.2 数据驱动的出版服务

在出版服务方面，CDL提供了开放学术出版平台eScholarship，平台提供了开放获取资源、学术出版研究工具，支持开放获取原始数字出版期刊、图书、论文等数据，也支持发表、检索、浏览、评论等操作。通过学术期刊、论文、图书等数据，智能化的平台工具，以及数据驱动服务，eScholarship为研究人员的创新提供了条件，为加州大学学者和研究人员提供了一个便捷、高效、低成本的出版方式，并支持了学术交流和学术出版

的可持续发展。

4.4 人文与个性化服务

4.4.1 数字驱动的人文个性化数字平台

CDL提供了多元而又丰富的数据资源，其中一些具有特色、较为人文化的数据被整理集成为多个特色平台。以档案数据为中心的OAC为用户提供了大量珍贵的历史文献和文化遗产，促进了历史和文化的研究，助推了知识的开放与共享，帮助用户了解和学习美国西部地区的历史和文化。专门收集政府机构发布的信息的CA.GOV则能够帮助研究者和历史学家找到过去的信息并进行进一步的相关研究。以文物数据为基础的Calisphere不仅提供了丰富的素材和信息，还促进了博物馆、图书馆和档案馆之间的合作和数字化进程，并帮助上述机构建设数据驱动服务。

4.4.2 数字驱动的实体书购买服务

提供数据驱动的实体书购买服务是CDL的个性化特色服务之一，为数据共享提供了新的思路。数字图书馆中的书籍、文献等资源均来自实体书籍，通过数字化技术CDL将这些实体书籍转化为数字书籍。以数字书籍为基础，CDL又提供了200 000种可获取纸质版书籍的重印本，读者可以通过亚马逊平台购买。虽然在数字化技术出现之前，书店也能提供纸质版书籍购买服务，但CDL提供的实体书购买服务是基于数据驱动工具的。通过数据驱动的实体书购买服务，将能够为用户提供更加丰富的实体书籍资源，满足更多用户的需求。

5 CDL数据驱动战略的启示

当前，学术图书馆已不再仅用于存储文献、书籍，而是能够精准对接用户需求，为用户提供真正需要的产品或服务[38]。CDL在学术型数字图书馆建设中具有引领地位，通过剖析CDL的服务转变方式，发现CDL具有以下优势：丰富的图书馆数据馆藏、智能化的数据驱动工具、较为全面的数据驱动服务方式，因此可以实现最大限度地发现和访问信息资源，优化和开发共享服务以提高运营效率，深度参与学术交流，加强图书馆服务、资源和运营的多样性、公平性、包容性和归属感。根据上述优势，得出以下启示，可供我国学术型数字图书馆参考学习。

5.1 构建数据管理流程，开发数据驱动工具

CDL通过丰富而又多样的馆藏以及智能、多样化的数据驱动工具实现了数字驱动的服务。CDL注重软件、工具与平台的自主开发，其开发成果备受关注，如Dash曾获得专业鉴定机构的认可、Merritt曾获相关领域奖项。这些自主开发的软件、工具使CDL长期立于领域内的主导地位。目前我国学术图书馆已然拥有大量的馆藏数据，但还未实现数据的循环，在数据的收集、标识、处理、管理以及数据驱动工具、数据驱动服务方面还有进步的空间。例如，缺少统一的数据描述与标识工具，缺少统一的馆藏数据决策与管理方式，缺少能够为科研工作者提供指导的数据管理计划工具等。全面建设数据驱动工具将能为数据的采集、描述、发布等提供保障，并进一步提升数据驱动的智慧服务能力。

5.2 推动科学交流，嵌入科研一线

从帮助科研人员了解数据管理规定信息和具体要求开始，到科研数据的发布、科研数据与文献的管理，再到科研数据的重新利用，CDL均提供了相应的数据基础、数据驱动工具以及数据驱动的智慧服务，通过数据驱动服务的方式，实现图书馆与用户、用户与用户之间的双向资源交流，将图书馆从文献搜索平台逐渐转变为学术交流的中心，让图书馆在学术交流的过程中变得越来越重要。目前我国图书馆通常以资源提供者的形式存在，图书馆与用户以及用户与用户之间的数据交流通道较少（尤其是用户与用户之间的数据交流通道）。通过建立更多的数据交流通道，嵌入科研一线，将能够帮助推动科学交流。

5.3 关注多形态数据建设，建立多元合作渠道

多元化的合作也是CDL的一大亮点。CDL的合作是多方面、多元化的，主要体现在馆藏的收集与开放获取的实现、数据驱动工具的建设等过程中。CDL的合作机构众多，包括高校、政府、企业等，通过与上述机构合作，CDL收录了更多元、丰富的数据。CDL不仅将加州大学十大分校的资源结合在一起，还与加州大学以

外的加州各大图书馆、博物馆、档案馆、历史馆等机构以及各大期刊，甚至全美资源库联合在一起。这些资源不仅丰富了学术界和研究领域的资源库，为教育和公共利益作出了重要贡献，还使CDL成为商业供应商的学术制衡力量。目前，我国图书馆主要聚焦文献、图书、档案、科学数据等资源的建设，还缺少对政府资源、视频资源等其他形态数据的关注，合作对象也较为单一，因此通过建立多元合作渠道能够促进多形态数据的建设。

参考文献

[1] TOLLE K M，TANSLEY D S W，HEY A J G. The fourth paradigm：data-intensive scientific discovery[J]. Proceedings of the IEEE，2011，99（8）：1334-1337.

[2] 申静，杨家鑫. 数据驱动的智库知识服务流程优化[J]. 图书情报知识，2021，38（4）：114-124.

[3] MANYIKA J，CHUI M，BROWN B，et al. Big data：the next frontier for innovation，competition，and productivity[EB/OL]. [2023-10-23]. https://www.mckinsey.com/business-functions/mckinseydigital/our-insights/big-data-the-next-frontier-for-innovation.

[4] CURUKSU J D. Data Driven：An Introduction to Management Consulting in the 21st Century[M]. Cham：Springer International Publishing，2018：20-21.

[5] 李洁. 数据驱动下数字图书馆知识发现服务创新模式与策略研究[D]. 长春：吉林大学，2020.

[6] CHEN C L P，ZHANG C Y. Data-intensive applications，challenges，techniques and technologies：a survey on big data[J]. Information Sciences，2014，275：314-347.

[7] 刘泽，邵波，王怡. 数据驱动下图书馆智慧参考咨询服务模式研究[J]. 情报理论与实践，2023，46（5）：176-184.

[8] CDL. About CDL[EB/OL]. [2023-11-22]. https://cdlib.org/about/.

[9] File-Analyzer[EB/OL]. [2023-08-23]. https://github.com/CDLUC3/File-Analyzer.

[10] UCLA Digital Library Program & UC3 collaborate to reaffirm a preservation solution for the Strachwitz Frontera Collection[EB/OL]. [2023-08-23]. https://uc3.cdlib.org/2021/11/11/ucla-dlp-uc3-collaborate-on-preservation-solution-for-frontera/.

[11] JOHN C. California Digital Library：advancing the digital transition of scholarly information[J]. Abstracts of Papers of the American Chemical Society，2016，251.

[12] 孙华. 基于馆藏结构优化的读者决策采购策略研究[J]. 新世纪图书馆，2023（1）：37-43.

[13] CDL. Mass digitization[EB/OL]. [2023-08-21]. https://cdlib.org/services/pad/massdig/.

[14] Contribute content[EB/OL]. [2023-11-22]. https://www.hathitrust.org/ingest.

[15] Dryad. Dryad partnering with CDL to accelerate data publishing[EB/OL]. [2023-08-23]. https://blog.datadryad.org/2018/05/30/dryad-partnering-with-cdl-to-accelerate-data-publishing/.

[16] Dryad. Who we are[EB/OL]. [2023-11-22]. https://datadryad.org/stash/about.

[17] University of California. eScholarship[EB/OL]. [2023-11-22]. https://escholarship.org/.

[18] 黄如花，李楠. 加州大学伯克利分校图书馆科研支撑服务研究[J]. 图书馆建设，2016（5）：46-50.

[19] 常进. 数据管理体系中学术图书馆的功能探讨：以美国加州大学数字图书馆为例[J]. 大学图书情报学刊，2014，32（5）：7-9，29.

[20] CDL. eScholarship[EB/OL]. [2023-11-22]. https://cdlib.org/services/pad/escholarship/.

[21] eScholarship. Open access publications from the University of California[EB/OL]. [2023-11-22]. https://escholarship.org/.

[22] STEPHEN A. Merritt and Dash certified as trustworthy repositories[EB/OL]. [2023-08-23]. https://uc3.cdlib.org/2018/08/22/merritt-and-dash-certified-as-trustworthy-repositories/.

[23] ERIC L. Merritt renews its CoreTrustSeal certification[EB/OL]. [2023-08-23]. https://uc3.cdlib.org/2023/04/05/merritt-renews-its-coretrustseal-certification/.

[24] ERIC L. Merritt is awarded a Mojgan Amini Operational Excellence Award at UC Tech 2022[EB/OL]. [2023-08-23]. https://uc3.cdlib.org/2022/08/19/merritt-is-awarded-a-mojgan-amini-operational-excellence-award-at-uc-tech-2022/.

[25] STARR J. EZID：a digital library data management service[M]//BAKER D，EVANS W. A Handbook of Digital Library Economics. Amsterdam：Elsevier，2013：175-183.

[26] RONALD C J. A report on the DataCite summer 2013 meeting[J]. Library Hi Tech News Incorporating Online & Cd Notes，2014，1（31）：4-7.

[27] EZID. Learn About EZID[EB/OL]. [2023-07-30]. https://ezid.cdlib.org/learn/.

[28] 黄如花，林焱. 加州大学伯克利分校数据管理的实践剖析[J]. 图

书情报工作，2016，60（3）：26-31.

[29] JACOB M. Data management plan：a background report[EB/OL]. [2023-07-30]. https://bdes.datasociety.net/wp-content/uploads/2016/10/DMPReport.pdf.

[30] CDL. Create Data Management Plans that meet requirements and promote your research[EB/OL]. [2023-11-22]. https://dmptool.org/.

[31] CDL. Mass Digitization of UC Library collections[EB/OL]. [2023-11-22]. https://cdlib.org/services/pad/massdig/.

[32] OAC. Welcome to the Online Archive of California[EB/OL]. [2023-11-22]. http://www.oac.cdlib.org/.

[33] CDL. CA.GOV Web Archive[EB/OL]. [2023-11-22]. https://cdlib.org/services/pad/webarchiving/ca-gov-web-archive/.

[34] Calisphere. The deeper you look，the more you discover[EB/OL]. [2023-11-22]. https://calisphere.org/.

[35] 吴敏. 美国加州大学图书馆战略规划分析与启示[J]. 数字图书馆论坛，2020（3）：66-72.

[36] UC Libraries. Systemwide ILS（SILS）[EB/OL]. [2023-11-28]. https://libraries.universityofcalifornia.edu/sils/.

[37] UC. Participate in the UC open access policies[EB/OL]. [2023-07-25]. https://osc.universityofcalifornia.edu/for-authors/open-access-policy/.

[38] 肖希明，尹彦力. 服务于“双一流”建设的高校图书馆信息资源建设[J]. 图书馆建设，2018（4）：79-84.

作者简介

王茜，女，博士研究生，研究方向：情报理论与方法。
孙蒙鸽，女，博士研究生，研究方向：情报理论与方法。
郑新曼，女，博士研究生，研究方向：情报理论与方法。
刘细文，男，博士，研究员，通信作者，研究方向：信息资源管理，E-mail：liuxw@mail.las.ac.cn。

Practices and Enlightenment on Data-Driven Strategy of California Digital Library

WANG Xi[1,2]　SUN MengGe[1,2]　ZHENG XinMan[1,2]　LIU XiWen[1,2]
(1. National Science Library, Chinese Academy of Sciences, Beijing 100190, P. R. China; 2. Department of Information Resources Management, University of Chinese Academy of Sciences, Beijing 100190, P. R. China)

Abstract: Domestic academic libraries are in the period of transformation to data-driven service. Taking California Digital Library (CDL) as an example, this paper analyzes the data-driven management framework and practices, and makes detailed investigation and analysis on the development of its data-driven tools, data-based infrastructure construction, and data-driven services. By sorting out the advantages of CDL, we propose three revelations, including strengthening the construction of data management processes and the development of data-driven tools, embedding in the frontline of scientific research and further promoting scholarly communication, and focusing on the construction of polymorphic data and establishing diversified co-operation channels. We hope to provide a reference for the transformation of academic library to data-driven services of Chinese academic libraries.

Keywords: Data-Driven Strategy; Digital Library; Library Service; University of California; Academic Library

（责任编辑：王玮）