大数据时代开放信息资源的数据服务能力思考

黄金霞，马雨萌
（中国科学院文献情报中心，北京 100190）

摘要：面对科研第四范式下科研人员对科学数据及其应用的需求变化，图书馆开始思考开放资源建设的服务能力。分析开放信息资源的数据化发展特征，基于对中国科学院科研人员数据需求的调查分析，设计开放信息资源的数据服务策略和服务流程，并初步开展数据服务实践。从为用户提供更精准的开放资源服务角度，建议图书馆深入研究数据服务的理论和方法。

关键词：科研第四范式；开放资源；数据服务；服务策略
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2016.8.009

近十年，在期刊订购经费危机、信息获取许可危机、学术交流开放危机的刺激下，开放获取运动发展迅猛，产生的直接结果之一是形成大量开放信息资源（下文简称为“开放资源”）。如何利用这些快速发展的开放资源，一方面，在经费不增加投入的情况下继续充实馆藏；另一方面，突破传统的数据库访问服务方式而实现内容服务，成为研究型图书馆面临的机遇和挑战[1]。开放获取（Open Access），自产生就具有免费获取、随时随地通过网络访问、永久获取、使用权广泛等主要特征，使图书馆基于开放资源的信息服务方式更加丰富。本文将思考图书馆开放资源建设的新型服务能力，分析数据服务能力，尝试构建图书馆信息资源精准服务的新型实现方式，期待大数据时代的图书馆在搜索引擎、大型网络公司的重围中赢得继续发展的机会。

1 大数据环境中开放资源建设面临的挑战和机遇

1.1 大数据环境中科研工作信息资源需求变化

2011年5月，麦肯锡公司发布《大数据：下一个创新、竞争力和生产力前沿》，提出“大数据”将引发新一轮的生产力增长与创新，成为竞争的关键[2]。其后，“大数据”迅速成为热点，在助力智慧城市建设、互联网金融发展、电子商务、社会安全等方面的应用价值日益凸显[3]。在科学研究领域，也形成“大数据”这种新的科学基础设施，推动科学研究走向数据密集型的第四范式，数据不仅是科学研究的结果，更是科学研究的新基础[4]。

科研信息需求从传统的科研文献信息向科学数据转变，越来越多的科学研究工作无需从头开始，而是建立在对现有研究数据资源组织、解析及利用的基础上。从不同领域的科研论文看，科研第四范式中的数据涵盖大数据集和非大数据集，并非仅是大数据集。在生物学领域大量DNA序列数据的挖掘用于生物信息学研究[5]，遥感大数据自动分析和数据挖掘成为遥感领域科研发展方向[6]；同时，PubMed来源的百万篇文章的图、表被用来挖掘和构建脑成像数据库[7]，Web of Science中千篇文章的摘要被分析后用以重新发表论文；前两个例子用到了大数据集，后两个例子利用了文献来源的非大数据集。

当前科研论文和原始科研数据还不能广泛地开放获取，存在严重的“数据鸿沟”，也有较大比例科学家不愿意发布或共享自己的试验数据。文献中已发布数据或开放资源，成为新的数据分析来源，包括文献资源中期刊论文的图表和开放数据（开放的政府数据、天气数

据、交通数据、网络中免费的社会经济新闻、市场分析数据、标准、专利等）。开放资源的开放性、关联性、知识化使其更容易被发现、获取、再利用。

1.2 开放资源建设的服务能力思考

开放资源在学术活动中已越来越重要，其数量和学术影响力快速上升。2015年，开放获取期刊数量超过1.2万种，PubMed Central（PMC）存储的开放论文超过370万篇，世界银行仓储中科研产品达到19 474份，Dryad包括10 274个数据包、32 979个数据文件，OpenAire更是由欧洲38个合作国共同开发的开放知识库网络。同时，大量的文献资源、网络信息资源被Google、美国国会图书馆、开放地理信息联盟等以开放关联数据、通用数据格式对外发布。开放信息也在迅速成为可计算的开放知识，开放获取本身就具有较清晰的使用权益申明，其中Libre Open Access允许对信息的复用，包括数据挖掘、作品衍生等。数据分析和处理技术在科学研究领域的广泛应用，将支持开放资源被更好地再利用[8]。

除着力构建开放资源集成内容体系，开放资源建设更要打造开放资源服务体系。开放资源服务体系，包括普遍服务内容，例如开放资源元数据集成服务平台、进行多源开放资源内容关联、提供定题集成、编辑虚拟期刊、建立开放资源评价服务等；也包括再利用服务内容，例如开放资源内容以关联数据方式重新发布、基于开放科研数据挖掘新知识、打造特定领域的开放知识环境等。综上所述，从信息资源组织角度看，开放资源的服务可以划分为3个层次及对应多个服务产品，见表1。黄永文等提出开放资源的6种再利用服务模式，包括集成检索服务、学术引用网络分析服务等，主要指信息层面的服务[9]；知识层面的服务指围绕知识概念和关系的发现；数据层面的服务指提供各种类型数据及其应用的服务。

表 1 开放资源的服务层次

开放资源的服务层次	开放资源类型	开放资源服务产品
信息层面的组织和服务	● 开放获取期刊 ● 开放获取图书 ● 开放课件 ● 开放会议资源 ● 开放科技报告 ● 开放学位论文 ● 开放社会经济信息等	● 单类型开放资源集成服务系统 ● 综合开放资源集成服务系统 ● 特定领域的虚拟期刊 ● 开放资源评价服务
知识层面的组织和服务	● 知识概念 ● 知识图谱 ● 知识发现	● 知识发现或学术关系发现系统 ● 知识问答系统 ● 知识环境
数据层面的组织和服务	● 元数据 ● 全文来源数据（图、表、数据等）和原始科研数据等 ● 各类知识组织体系	● 开放资源的开放数据接口 ● 开放资源的数据定制 ● 多种类型数据挖掘和分析产品

在数据密集型科研中，信息共享、交流互动已不再是最迫切的用户需求，数据的分析和整合才是最大挑战，因为数据是信息、知识和智慧的“原材料”[10]。图书馆亟需探索新的服务模式为用户提供精准服务，开放资源的快速发展，为图书馆开创数据密集型科研的个性化数据服务提供可能，但也面临理论和方法的挑战。本文思考的数据服务不等同于科学数据管理服务。科学数据管理服务是指为科学家提供科研过程中数据管理方案和存储服务，而开放资源的数据服务指提供来自于大量开放资源的数据获取、组织、分析等增值服务。当前国外图书馆提供的数据服务主要是数据分析服务，例如，美国康奈尔大学图书馆提供大型数据集的数据统计分析、空间分析、定性分析等[11]，这些服务内容并不包括数据获取服务、数据组织服务，因为可能牵涉到数据权、使用权等问题。当前绝大多数开放资源执行的知识共享协议可以帮助图书馆规避数据服务中的一些约束，例如，再利用权益问题[12]。利用开放资源进行数据服务的案例，目前在国内图书馆界并不多，绝大

多数图书馆缺乏大数据时代下的数据储备能力、数据服务能力，以及数据再利用的复杂权益问题处理能力。在开放资源集成建设的基础上，少数研究型图书馆正在尝试开展数据服务，例如开放论文一站式发现平台GoOA提供开放数据定制服务[13]。

2 开放资源的数据化发展特征

科研第四范式中使用的数据，包括大数据集和非大数据集。非大数据集的数据特征，包括中小规模、非时变、单一结构/领域、集中存储；而大数据集的数据特征，包括海量、分布/多源、流数据、异构、高不确定性等[14]。综合二者，信息资源的数据化发展特征，应包括规模化、多样化、结构化、价值化。

开放资源尤其是开放学术资源，例如开放获取期刊、开放获取图书、开放学位论文、开放课件等；生产、传播具有与传统文献资源基本一致的流程，包括编辑（出版）、交流、存储、再利用。开放资源出现最初是为了被广泛地发现和使用，在其产生和传播的不同阶段都为此目标作准备，包括数据化发展方向。

在编辑出版阶段，很多开放资源始于数字，资源内容越来越多以语义增强出版、结构化或半结构化的方式呈现，除资源本身规模的快速增加，OA论文、开放图书、开放报告、开放学位论文等开放资源中附带的开放数据、附录资料数据等也越来越多；在交流阶段，传统的文献资源或馆藏内容，被图书馆或其他信息建设机构逐步加工成开放关联数据，对外发布、开放共享，这也增加了开放资源的类型、规模和结构化程度；在存储阶段，中国科学院机构知识库网格集中存储的资源类型超过10种、资源数量超过70万个文件、来源于100多个研究所机构知识库，开放学科存储PMC和预印本系统arXiv收录的OA论文、数据仓储Dryad收录的数据文件已成规模，其中高价值的OA论文和数据不断被发现和应用，例如PMC的论文正被不同研究目标的科研人员进行挖掘分析；再利用阶段，数据作为一种公开商品或资源，早就明码标价以购买版权使用，而开放资源具有较清晰的使用权益，使其数据使用和再利用也更方便。

数据的价值化指在大数据的分析中对数据去冗分类、去粗取精，从数据中挖掘出有价值的信息与知识，把大数据通过定量分析变成小数据的过程[15]。来源于开放学术资源的科学数据，其质量和价值已经在同行评议和发表中被证明。

3 中国科学院科研人员的开放数据服务需求调研

中国科学院科研人员自2013年从事开放资源建设工作起，每年都在中国科学院进行用户需求问卷调查，调查结果反映科研人员对开放资源的需求变化：2013年，用户希望建设开放资源的发现途径、集成检索；2014年，用户希望提供开放资源的全文集成和获取方式；2015年，用户希望建立开放资源的发现和整合工具；2016年，科研人员对存在于不同开放信息源中的数据需求变得越来越强烈，包括对多源数据的发现需求、数据获取需求、数据整合需求、数据分析服务需求、数据挖掘服务需求等。因此，2016年3月组织的中国科学院科研人员开放数据需求问卷调查，目的在于掌握科研人员在数据利用过程中的问题和需求，为图书馆顺利开展基于开放资源的数据服务提供依据。

本次网络问卷调查，共收到反馈640份，包括来自中国科学院65个研究所的科研人员，其中，生物领域人员占31%，物理领域人员占19%，生态/环境领域人员占19%，化学领域人员占7%，计算机科学/自动化领域人员占5%[16]。问卷调查内容分为5类，如表2所示。

表 2 中国科学院科研人员的数据需求问卷调查

调查项	调查者反馈结果
对数据的兴趣	71%的调查者知道可以利用数据挖掘或分析方法来发表论文；95%的调查者对这种研究方法感兴趣
科研中常用的数据类型	80%的调查者认为在科研过程中最常用的数据类型为文献中的数据和原始试验数据，其次是开放网络信息中的数据，占46%；社会调查数据和商业数据，分别占26%和16%；有10%的调查者表示不需要大量数据集
数据获取方式	84%调查者以人工方式下载和积累数据；30%调查者通过付费方式寻求信息服务机构帮助获取数据或从商业公司购买数据，10%调查者认为费用过高；26%调查者在感觉困难时放弃获取数据

续表

调查项	调查者反馈结果
数据利用中遇到的困难	58%调查者不知道从哪里获得所需要的数据，且无法把握数据的准确性和可靠性；60%调查者不知道如何整理数据，且缺乏可用的数据采集工具、处理工具和分析工具；36%调查者找不到合适的机构进行数据服务咨询；10%调查者提出与IT公司的技术人员沟通不畅，无法很好地解决数据利用问题
希望获得的数据服务方式	78%调查者期望与图书馆这样的非营利性机构进行合作；66%调查者希望获得数据整理加工的定制服务；50%调查者希望获得数据挖掘和分析服务；47%调查者希望获得数据利用咨询服务；超过50%调查者希望图书馆提供相关技术培训、支持工具

从表2的调查结果可以看出，在大数据时代的茫茫数据海洋中，科研人员有强烈的数据使用需要，但目前还无法有效地发现数据、获取数据和应用数据，期望能获得非营利性的定制数据服务渠道。另外，数据服务需求偏向于生物医药、地球物理等相关领域。

4 基于开放资源的数据服务策略和流程设计

依据上述对开放资源的数据化发展特征分析以及用户需求分析，参照《数据工程理论与技术》[10]，设计基于开放资源的数据服务策略及工作流程（见图1）。

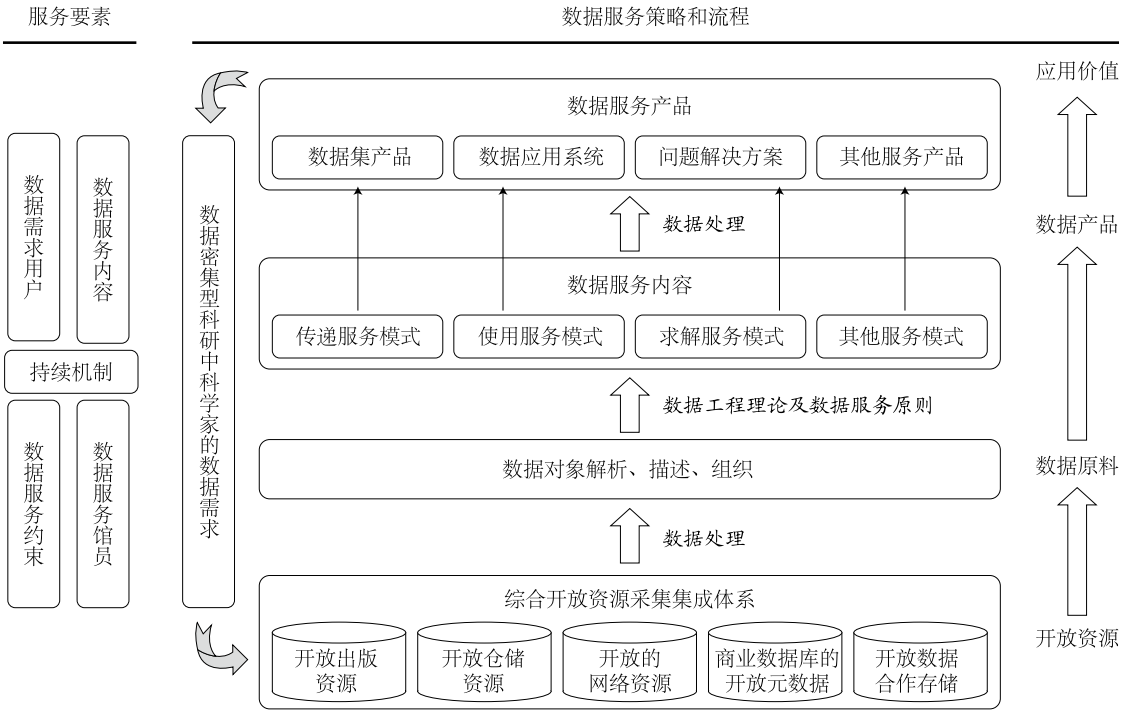

图 1 基于开放资源的数据服务策略和流程设计示意图

4.1 数据服务策略设计

其主要包括服务要素构建策略、服务产品构建策略和服务持续策略。

（1）数据服务要素构建策略。与数据工程服务或图书馆的传统文献服务不同，开放资源的数据服务是面向科学家的个性化研究需要，数据类型涉及面广，需要针对用户需求快速设计出有效的数据服务内容，需要强有力的技术支持，有知识产权问题或权益纠纷处理能力。所以，要构建研究数据服务要素以保障服务的顺利开展。数据服务要素包括数据需求对象、数据服务内容、数据服务约束、数据服务馆员的构建，及其在服务策略和流程中的相互作用和彼此关系。数据服务馆员团队的培养很重要，其需要具备开放资源的发现获取

能力、数据化处理能力、数据服务内容设计能力、数据服务产品实现和应用能力，包括学科馆员、资源建设人员、计算机专业的技术人员等。

（2）数据服务产品构建策略。依据用户数据服务需求，研究设计多种类型的数据服务模式，包括多种数据服务内容及其服务产品设计。按照数据工程理论，数据服务模式按照层次高低可设计为传递服务模式、使用服务模式、求解服务模式及其他服务模式，所对应的数据服务产品相对应为数据集产品、数据应用系统、问题解决方案及其他服务产品[17]。数据服务有个性化特征，有必要按照用户的特定需求来构建服务产品，同时，服务内容不同，其涵盖的数据服务要素也将发生相应变化。

（3）数据服务持续策略。图书馆开展开放资源建设，对用户来说并不比Google或百度等开放搜索引擎的吸引力大，但当图书馆拥有开放资源内容时，将开创个性化的数据服务方式，这将为图书馆资源建设工作提供新的发展方向。所以有必要围绕数据服务能力和服务流程来研究建立服务可持续机制，包括数据服务各要素的能力持续建设、服务策略持续建设，以及服务过程中的约束分析和管理方案建设等。

4.2 数据服务流程设计

以上服务策略将保障从开放资源发现到数据原料加工，从数据产品构建到应用价值实现的服务流程。流程具体包括：了解数据密集型科研中不同领域科研人员的数据需求，例如数据发现、数据获取、数据加工、数据分析等需求；目标开放资源的发现和采集，开放资源范围涵盖用户所需要的类型，例如开放出版资源、开放仓储资源、开放的网络信息、商业数据库的开放元数据、开放或合作的科研数据等；信息资源的数据化解析和组织，完成数据格式清理、内容清理，建立一定程度的数据关系，例如水稻品种数据与表型数据的关系、表型数据与基因突变体DNA片段数据的关系；数据服务内容设计，构建用户需要的数据服务模式及其服务产品形式，并协助用户进行进一步的数据分析和论文撰写等。

4.3 开放资源的数据服务实践

多年来国内水稻的品种改良工作一直在进行，相关研究人员发表了众多论文，尤其是在提升水稻抗性方面的工作进展很大，但很多工作是通过常规育种方式进行的，如何把国内的常规育种结果与国外分子生物学层次的研究结果结合起来进行系统分析，目前国内的科研人员还缺乏有效的平台和工具。水稻分子育种的科研人员希望从多源开放资源中进行相关数据获取、加工和分析，以建立可用的数据集，再利用这个数据集进行相关数据分析，推进科研试验，撰写文章并发表。

中国科学院科研人员面对水稻开放数据服务需求，首先，确定由生物领域学科馆员、开放资源建设人员、数据加工技术人员组成的3人数据服务小组；然后与科研人员确定需求的数据类型、数据获取源、数据服务模式；最后，确定服务完成时间和后续数据更新保障时间（年）。目前服务产品已完成，产品形式为数据集文件，数据来源和类型涉及开放网络来源的国内水稻审定品种、亲本、性状尤其是具有的抗性，开放文献来源的水稻品种/品系、突变体基因型、对应的表型、突变基因片段的PRC引物序列及其PCR电泳图谱。包括50年来的8 000多个水稻省级审定品种的亲本信息、27个表型特征，例如茎、叶、穗、粒、植株、种子、敏感度、米质等的性状；30多种抗性例如稻瘟病抗性、白叶枯病抗性等及其抗性等级；10条基因描述信息例如基因座名称、所在染色体、定位与克隆、突变体表型等，后续数据更新时间为1年/次。

5 结语

当大数据浪潮扑面而来的时候，图书馆这只海燕该如何飞翔[18-19]。现有的大数据应用多集中于基础设施建设（云平台、数据中心、计算架构等），所展现的成功应用基本是以查询处理为基础的技术，分析还仅限于传统方法（统计分析、数据挖掘），仍是非常初级的[14]。同时，在科研领域的数据分析，使用到大数据集和非大数据集，要发展大数据产业，数据依然是基础，选择合适粒度的数据，集成这些数据，需要重视数据资源建设，但当前开放数据资源建设仍存在一定资源获取难度。

数据密集型科研用户对数据的强烈需求，为资源建设方式的转型提供目标，开放资源再利用建设将可能“从数据直接实现价值”，而不再依赖传统的信息传递链。图书馆可以考虑抓住开放资源的快速发展机会，紧密结合科研用户对数据的需求，培养数据服务馆员，深度打造一批数据资源，为用户提供定制性的服务产品，形成本馆在大数据发展时代的精准服务能力之一。研究型图书馆与科研人员的贴近，具备传统资源发现、信

息组织、采集技术和情报分析能力，使图书馆开展个性化数据服务成为可能，当然，图书馆还需要在数据服务理论和方法上进行系统而深入地研究。

参考文献

[1] 黄金霞,张建勇,黄永文,等.开放资源建设的措施及工作策略[J].图书情报工作,2013,57(8):57-61.

[2] McKinsey Global Institute.Big data:the next frontier for innovation,competition,and productivity[EB/OL].(2011-05)[2016-07-23]. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.

[3] 李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J].中国科学院院刊,2012,27(6):647-657.

[4] 朱扬勇,熊赟.DNA序列数据挖掘技术[J].软件学报,2007,18(11):2766-2781.

[5] 李德仁,张良培,夏桂松.遥感大数据自动分析与数据挖掘[J].测绘学报,2014,43(12):1211-1216.

[6] HEY T,TANSLEY S,TOLLE K.第四范式:数据密集型科学发现[M].潘教峰,张晓林,译.北京:科学出版社, 2012.

[7] YARKONI T,POLDRACK R A,NICHOLS T E,et al.Large-scale automated synthesis of human functional neuroimaging data[J].Nature Methods, 2011,8(8):665-670.

[8] 张晓林.开放获取、开放知识、开放创新推动开放知识服务模式——3O汇聚于研究图书馆范式再转变[J].现代图书情报技术,2013(2):1-10.

[9] 黄永文,张建勇,谢靖,等.开放资源的再利用模式研究[J].图书情报工作,2013(21):32-37.

[10] 戴剑伟,吴照林,朱明东,等.数据工程理论与技术[M].北京:国防工业出版社, 2010.

[11] Cornell University.Data management services at Cornell[EB/OL].[2016-07-23]. http://data.research.cornell.edu/services#Data collection and analysis.

[12] 刘静羽,肖曼,陈雪飞,等.图书馆开放期刊再利用中的权益问题研究[J].数字图书馆论坛,2016(2):63-71.

[13] 用GoOA数据,做你的文章——GoOA文献数据再利用服务[EB/OL].[2016-07-23]. http://gooa.las.ac.cn/external/open_interface_datause.jsp.

[14] 徐宗本.大数据·大智慧——“大众创业、万众创新”背景下的大数据产业[EB/OL].(2016-01-11)[2016-07-23].http://www.casmooc.cn/onlineCourseAction.do?method=detail&bookId=1452478828623.

[15] 李广建,化柏林.大数据分析与情报分析关系辨析[J].中国图书馆学报,2014(5):14-22.

[16] 开放资源建设团队.中国科学院科研人员的开放数据需求调查报告[EB/OL].[2016-07-23].http://ir.las.ac.cn/handle/12502/8706.

[17] 文峰.对几种典型数据服务模式的对比分析[J].科技信息, 2013(25):107-108,139.

[18] 苏新宁.大数据时代数字图书馆面临的机遇和挑战[J].中国图书馆学报, 2015 (6):4-12.

[19] 张斌,马费成.大数据环境下数字信息资源服务创新[J].情报理论与实践, 2014(6):28-33.

作者简介

黄金霞，女，1972年生，博士，中国科学院文献情报中心副研究馆员，研究方向：信息资源组织和建设、开放资源建设，E-mail:huangjx@mail.las.ac.cn。
马雨萌，女，1989年生，硕士，中国科学院文献情报中心助理馆员，研究方向：开放资源建设。

Thinking on the Data Services from Open Resources Development in Big Data Era

HUANG JinXia, MA YuMeng
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Based on the demand changes to scientific data and its applications from the researchers in the fourth paradigm, the service abilities in the open resources development were studied. Firstly, the data characteristics of open resources were analyzed, and the demands on scientific data were also surveyed in the researchers of Chinese Academy of Sciences. Then, this paper designed the data service strategy and service process of the open resources, and completed a data service practice. From the perspective of providing researchers with more accurate open resource services, it was discussed finally that the library should pay more attentions on the theory and methods of data services.

Keywords: the Fourth Paradigm; Open Resources; Data Services; Service Strategy

（收稿日期：2016-08-23）