

不同科研场景对数字信息资源的依赖性分析

于倩倩¹ 贾茹^{1,2} 黄金霞¹

(1.中国科学院文献情报中心, 北京 100190; 2.中国科学院大学, 北京 100049)

摘要: 文章分析了新型学术交流模式、e-Science、数据密集型科研等不同科研场景对数字信息资源的依赖, 并提出应对策略。新型学术交流模式主要依赖于数字资源的开放关联、组织描述、深度开放以及免费共享; e-Science主要依赖于数字资源的元数据描述、内容存储关联等多种服务, 并且资源依赖具有学科特征; 数据密集型科研主要依赖于海量科学数据资源以及多类型软件工具类资源等。针对学术交流模式的变革提出信息资源的多样化建设尤其是开放科学数据的建设, 并嵌入到学术交流过程中; 针对e-Science的发展, 提出把开放信息资源建设嵌入到科研环境中, 关注不同学科对信息资源的需求特点; 针对数据密集型科研, 提出加强工具类资源的建设, 提供与数据相关的更多服务等。

关键词: 学术交流; e-Science; 基础设施; 数据密集型科研; 第四范式

中图分类号: G255

DOI: 10.3772/j.issn.1673—2286.2014.05.003

1 引言

在开放数字环境下, 从科研网络模式、知识库模式、开放获取模式等为标志的新型学术交流模式的出现, 到e-Science和数据基础设施为标志的科研环境的变化, 再进入数据密集型科研即第四范式, 科学研究本身发生了很大的变化。这些科研场景依赖于先进技术和大量信息资源, 已经嵌入科研工作流的不同环节, 也逐步走向基础设施化。它们的变化形成对相关信息资源的数字化、开放性、数量、类型等的需求变化, 而开放数字信息资源在知识组织、数据挖掘、可视化、设施化等方面的“信息服务速率”也影响着科研效率。这些多元化的信息资源需求和利用方式, 为图书馆资源建设带来了新的挑战。本文调研了国内外关于新型学术交流模式、e-Science、大型基础设施、科研第四范式等的研究报告、相关文献和网络信息, 以了解不同科研场景对数字信息资源(既包括数字文献资源、开放资源、网络资源, 也包括科学数据资源、软件工具类资源等)的需求和依赖, 希望能够把握新形势下对信息资源依赖的发展趋势。

2 新型学术交流模式对数字信息资源的依赖

新型学术交流模式是相对于以印刷出版的学术期

刊作为科研成果主要传递方式的传统学术交流模式而言, 其主要特点是以数字信息资源为中心。例如, 美国学者J. M. Hurd在Garvey-Griffith模型基础上提出的2020年学术交流模式^[1]; 作为网络环境下非正式交流有效形式包括科研语义网、网络论坛、科研博客等的科研网络模式; 以收集、描述、保存、传播机构成员所创造的数字化知识为主要内容的机构知识库模式; 为打破商业出版垄断而推动科研成果网络自由传播的开放获取模式; 着重于学术性质交流在虚拟环境下组建群组的虚拟学术群组模式等。

2.1 更多依赖于数字化资源

信息技术的发展使得研究记录从数据产生到稿件提交以至最后成果出版实现全面的数字化, 信息交流的媒介也实现从以纸质文献为中心到以数字文献为中心的转变。在2020年学术交流模式中, 交流的便利使得科学家从个体研究逐步转向合作研究。学者将更多地依赖数字文献而不再是纸质文献进行学术研究和交流。预印本服务器、电子期刊网站以及聚合体网站以及数字图书馆将取代期刊、会议录等纸质文献成为主要交流媒介。此外, 科研网络模式、机构知识库模式、开放获取模式、虚拟学术群组模式都主要依赖于数字资源进行学术知识的交流和共享。

2.2 依赖于数字资源的开放关联

以RDF、OWL等为基础的VIVO^[2]被视为科研网络的语义探索,它涵盖了康奈尔大学、印第安纳大学、佛罗里达大学等8所高校的科研人员数据,内容包括科研人员的名字、头衔、研究领域、所属机构、活动和产出、研究资助等,能够发现跨学科、跨地域的科学家的合作。Harvard Catalyst (简称“HC”)^[3]整合了哈佛大学、波士顿大学、塔夫斯大学等机构的各种资源、技术和最佳临床实践,帮助这些项目成员单位共享工具和技术,目前已发展成一个强大的搜索引擎,帮助科研人员把检索结果放到一个语境中,关联提供其他更多信息资源,来充分揭示HC所包含的工具和数据库,汇总研究人员所需要的有关人、出版物、设施等信息。

2.3 依赖于数字资源的描述和组织

网络化信息联盟(CNI)秋季成员会议于2012年12月10-11日在阿林顿召开,探索了如何管理、维护、共享、识别、保存、利用并构建各类型数字内容(重点是研究数据)的全球社区内容展示^[4]。机构知识库应针对所保存的知识内容,针对保存、传播、利用的多元需要,建立灵活的内容描述和组织机制,可根据组织架构、主题、项目等来组织内容,还应通过知识成果的各种属性(如作者、所属部门等)将相关成果灵活关联。如arXiv^[5]从不同学科如物理学、数学、计算机科学等组织文献,针对每篇文献,具有题名、作者、主题、摘要、全文获取格式等描述项,作者描述项链接到此作者发表的所有文献等。

2.4 依赖于数字资源的细粒度、深度开放

知识界及专业学术出版社协会指出未来的学术交流系统,应考虑以数据集、模拟、软件和动态知识的表述等为信息单元^[6]。政府大数据如Data.gov在CKAN上推出了一个新的数据目录,开放成为一种趋势^[7]。另外开放科学数据、研究数据及元数据、研究出版物、学术会议论文、开放知识等都是科研人员开放获取的需求。科学数据方面,Jim Gray认为互联网能把所有科学数据与文献联系在一起,创建一个数据和文献互操作的世界^[8],自然出版集团发布了科学数据开放获取平台^[9],在研究出版物方面,2012年10月SCOAP3^[10]正式启动,

2014年起实现高能物理领域高水平学术论文的开放出版等。

2.5 依赖于学科数字资源的免费共享

MIT开放课件平台^[11]、MOOC模式^[12]凭借其大量免费内容以及高水平的教学成为越来越受欢迎的在线学习选择,MIT开放课件是麻省理工学院公开发布的课程资料,不需要任何注册登记,不限学习人员。Emory大学图书馆计划创建数字学术共享空间(DiSC),支持人文学科和跨学科研究,还将建立支持健康与生命科学的信息学研究中心,提供专业知识^[13]。学术博客、学术论坛、社交性网络以用户自主创建的资源为主,别人围绕自己的话题进行讨论,形成一对多的交流。

3 e-Science对数字信息资源的依赖

2000年11月,John Taylor^[14]首先提出,e-Science是在重要科学领域中的全球性合作,以及使这种合作成为可能的下一代基础设施,可称之为e-Science的广义定义。学者雷枫^[15]认为,e-Science是在科研活动各环节采用先进信息技术提供支持与服务的环境和由此产生的新的科研模式,可称之为e-Science的狭义定义。基础设施是支持数据获取、存储、管理、集成、挖掘和可视化的包含硬件、软件、网络和中间件的综合系统。e-Science下的科研过程主要包括项目立项、项目实施、项目总结等阶段。数据生命周期是e-Science中的信息资源流,是指从数据产生,经数据加工和发布,最终实现数据再利用的一个循环过程^[16]。在e-science环境下,不同学科领域的发展不尽相同。

3.1 依赖于数字资源的元数据描述

欧盟信息化基础设施咨询工作组(e-IRG)发布的欧盟与科学数据管理相关活动的调研报告,对基础设施中的元数据和质量问题提出了建议,研究了元数据描述的基本原则和对存储在资源库中的资源的质量要求,这些原则和要求被认为是所有研究基础设施的基准。针对元数据报告提出科研人员亟需能够描述各种科研资源和服务的元数据,各学科越来越需要共同商定具体的语义元素,描述性元数据应当包括出处信息

以支持长期保存等^[17]。荷兰e-Science中心(NLeSC)希望能够通过基础设施促进合作研究,寻找更好的生物标记物,为此,项目将设计通用的元数据,以允许利用网络服务进行数据交换^[18]。

3.2 依赖于数字资源类型的多样化

在项目立项阶段,科技发展政策与战略、学科热点和前沿信息以及科技查新是科研人员的普遍需求;在项目实施阶段,助理研究员、研究生等希望能参加更多的资源与工具使用的培训活动;在项目总结阶段,需要查大量的文献解释研究结果,获取创新性成果^[19]。整个过程中,课题跟踪信息比较重要,而且用户需要的信息资源类型多样,除了图书、期刊、专利文献、会议文献,还包括数据库、多媒体资源、网络资源、实验数据、重要的规章制度、国内外研究计划、科研项目相关信息、相关领域或项目专家资料与联系方式、产品或资料等信息资源^[20]。

3.3 依赖于数字资源内容存储、关联等多种服务

在数据生命周期中,用户具有多种服务需求^[16]。例如,数据发现与获取服务(如CISTI提供数据导航服务)、数据存储服务(如普渡大学的D2C2)、数据长期保存服务、数据管理服务(如康奈尔大学数据管理中心)、数据关联服务(如生命科学搜索引擎Entrez,允许在读一篇文章时,链接到相关的基因数据,随着该基因到相关疾病的数据,再回到与这个疾病相关的文献^[8])、数据挖掘分析服务(如田纳西大学开展的数据分析咨询服务)等。基础设施也会根据科研人员需求,提供不同服务,例如NSF-ACP^[21]的Cyber-Infrastructure提供高性能计算服务、数据管理服务、接口、可视化服务、协同服务等。

3.4 数字资源依赖具有学科特征

e-science环境中不同学科领域用户对信息资源的需求具有“学科特征”。例如微生物学科领域需要全球性的资源合作共享,具体到我国微生物学领域,需要统一的、近乎覆盖全国的信息平台和知识环境,以及具有国际影响力的微生物学数据库^[22];海洋声学观测领域

的基础设施还缺乏,对基础硬件环境、数据处理软件系统以及了解海洋领域的信息技术人才都有需求^[23];野外观测活动中还无有效的信息资源移动服务支撑^[24];而同时地球科学领域已提出在e-Science环境下,信息资源内容及获取方式更加多样、存储大规模化、信息资源语义化,并提出数据挖掘和协同合作的资源需求^[25]。

4 数据密集型科研对数字信息资源的依赖

2007年,Jim Grey在NRC-CSTB的演讲报告中提出,数据爆炸和信息处理技术使得科研方法从之前的实验型科研(Experimental Science)、理论型科研(Theoretical Science)、计算型科研(Computational Science)推进到第四范式——数据密集型科研(Data-intensive Science)^[8]。数据密集型科研通过仪器收集或仿真计算产生数据,然后用软件进行处理,再将形成的信息和知识存储于计算机中,数据密集型科研是大数据背景下e-Science发展的新阶段。

4.1 依赖于多来源、大规模的数据资源

数据密集型科研场景下数据资源具有不同规模和性质的来源,如大型国际实验室、跨实验室、单一实验室或个人观察实验,还有可能来自于个人生活之中。海量的科学数据正迅速产生、广泛传播,例如美国的大型天文观察望远镜(LSST)投入运行后第1年所产生的数据就达到1.28PB,欧洲分子生物实验室核酸序列数据库(EMBL-Bank)近年收到数据的速度每年递增200%,在生物医学文献编目中已经有1800万篇医学文章,现在每年增加近100万篇。在数据密集型科研场景下,数据越来越重要,不仅可以作为科学研究的对象和工具,还可以基于数据进行思考、设计和实施科学研究。

4.2 依赖于多类型的软件工具类资源

各种实验涉及许多学科、大规模数据,特别是它们的高数据通量,使得合适的采集、管理和分析工具成为巨大的挑战。澳大利亚的平方公里阵列射电望远镜项目、欧洲粒子中心的大型强子对撞机、天文学领域的泛STARRS天体望远镜阵列等,每天都能产生好几个千万

亿字节的数据,但现在只能按照可管理的能力限制其数据速率。需要创建一系列通用的工具以支持从数据采集、验证到管理、分析和长期保存等整个流程^[8]。目前,无论在超大规模或微细规模的科学研究中,采集数据、管理数据、分析数据的工具都不能令人满意,在数据密集型科研环境下,亟需对海量数据进行管理、挖掘与可视化分析的新工具、新方法。

4.3 依赖于数据资源的元数据、语义描述

科学数据采集、处理、转换、转移过程,几乎没有可靠的元数据进行描述,对科学数据的权属、权利转让、管理要求、使用许可等更缺乏规范的机器可读的元数据,对于数据单元和数据集的标识和引用,缺乏广泛认可和可互操作的唯一标识符体系和引用规范,亟需建设描述科学数据的元数据。随着数据资源体量、复杂性和异构性的增加,科研人员越来越需要一些基于语义方法的新能力(如以知识本体方式来组织知识)。科学数据基础设施的开发者越来越需要基于语义的方法、工具和中间件,推动科学知识的建模,实现基于语义的数据集成,能够对不同系统中的数据实现集成化的知识发现和分析。

4.4 数据资源的依赖具有学科差异

不同学科领域在数据密集型科研中,对信息资源的需求是不同的。医疗领域对知识转化、信息资源共享和大规模处理工具有需求,生命科学领域对数据的挖

掘和能支持数据的基础设施有需求,地球与环境科学领域对相关工具和数据管理的需求比较强烈^[8],图书情报领域对开放数据获取具有需求,例如美国堪萨斯大学图书馆^[26]希望通过提高KU ScholarWorks等系统的元数据质量、更好地建立和促进对数据管理计划的支撑、投资开发新的数据发现工具和数据出版工具(例如汤森路透公司的数据引文索引)、推动开放数据运动并参与其中、开展研讨会和咨询活动等推进数据获取(共享、出版)以及提升数据素养。

5 不同科研场景下资源依赖比较分析及应对策略

新型学术交流模式主要依赖于数字资源的开放关联、描述和组织方式、细粒度深度开放以及免费共享; e-Science科研环境依赖于数字资源的元数据描述、多类型化、内容存储关联等多种服务,并且资源依赖具有学科特征;在数据密集型科研环境下,依赖于海量科学数据资源、多类型软件工具类资源、元数据及语义描述,并且资源依赖具有学科差异。三种科研场景都体现了对数字资源知识组织、开放共享、关联挖掘等方面的需求,但在具体的数字资源类型、知识组织方式、数据处理规模等方面存在差异,见表1。

5.1 新型学术交流模式下的应对策略

对于学术交流模式的变革,可以考虑的应对策略包括:(1)信息资源的多样化建设,着重在开放资源、

表1 不同科研场景下数字资源依赖比较分析

数字资源依赖 \ 科研场景	新型学术交流模式	e-Science	数据密集型科研
数字资源类型	主要是数字文献资源、开放资源、网络资源	主要是数字文献资源、开放资源、网络资源,部分科学数据资源	主要是科学数据资源以及工具类资源
数据规模	规模较小	规模中等	规模很大
知识组织方式	数字资源内容的元数据描述、语义关联	基础设施中资源元数据的描述	科学数据元数据描述、科学数据基础设施语义描述
服务方式	开放关联	存储、管理、关联、挖掘分析等	海量数据管理、分析、知识发现等
开放程度	细粒度、深度开放	协同共享	逐步开放
学科特征	学科特征不明显	学科特征明显	学科特征明显

用户自建资源,尤其是开放科学数据建设,提供用户参与信息资源建设的渠道;(2)加强信息资源的数据挖掘能力、信息组织能力、可视化能力的建设;(3)把资源建设工作嵌入到学术交流过程中,支撑不同环节对信息资源类型、组织方式、服务方式的需求。

5.2 e-Science下的应对策略

对于e-Science科研环境的变化,可以考虑的应对措施包括:(1)把开放信息资源建设工作真正嵌入到科研环境中,从数据生命周期来设计资源建设内容和服务内容;(2)关注不同学科发展中对信息资源的需求特点,加强资源合作共享能力;(3)尤其关注信息资源的知识组织方式,包括元数据的、语义的,以期能支持e-Science的数据设施化;(4)加强信息资源的服务内容,例如数据存储服务、数据管理咨询服务、资源移动服务等。

5.3 数据密集型科研下的应对策略

对于数据密集型科研的变化,可以考虑的应对措施包括:(1)提供与数据相关的更多内容建设、组织建设、服务建设,包括数据存储、计算、挖掘、可视化等;(2)要面向不同学科主题来组织、集成相关信息资源;(3)增加数据分析工具的建设;(4)通过推动开放获取运动、加强开放资源全文建设的方式,参与到未来的科研过程中,包括加强开放期刊建设、开放科学数据建设、国家仓储性质的知识库建设等。

6 结语

在开放数字环境下,支持科研创新、合作和交流的科研场景发生着变化,这对数字信息资源的建设和服务提出了新的挑战。图书馆作为重要的信息服务机构,必须把握新形势下科研场景的变迁,积极转变和拓展数字资源内容和服务方式。本文对新型学术交流模式、e-Science、数据密集型科研场景及其对数字信息资源的依赖进行了分析和比较,并针对其各自对数字信息资源依赖的特点,提出各个科研场景下信息资源建设的应对策略,希望能够为图书馆资源建设提供一些参考和借鉴。

参考文献

- [1] HURD J M. The transformation of scientific communication: a model for 2020 [J]. Journal of the American society for information science, 2000, 51(14): 1279-1283.
- [2] Cornell University Library. Vivo [EB/OL]. [2013-12-24]. <http://vivoweb.org>.
- [3] Harvard Catalyst [EB/OL]. [2013-12-10]. <http://catalyst.harvard.edu/>.
- [4] 2012 CNI Fall Membership Meeting: Scholarship for the Future [EB/OL]. [2013-01-20]. <http://www.dlib.org/dlib/january13/morris/01morris.html>.
- [5] Cornell University Library. arXiv.org [EB/OL]. [2013-01-20]. <http://arxiv.org/>.
- [6] 孙坦.开放信息环境:学术图书馆信息资源建设的重定义与再造[J].中国图书馆学报,2013(03):9-17.
- [7] What Governmental Big Data May Mean For Libraries [EB/OL]. [2013-12-14]. <http://lj.libraryjournal.com/2013/05/oa/what-governmental-big-data-may-mean-for-libraries/>.
- [8] Tony Hey,等.第四范式:数据密集型科学发现[M].潘教峰,等译.北京:科学出版社,2010.
- [9] NPG to launch Scientific Data to help scientists publish and reuse research data [EB/OL]. [2013-12-20]. http://www.nature.com/press_releases/scientificdata.html.
- [10] About SCOAP3 [EB/OL]. [2013-12-22]. <http://scoap3.org/about>.
- [11] MIT OpenCourseWare [EB/OL]. [2013-12-22]. <http://ocw.mit.edu/index.htm>.
- [12] MOOC [EB/OL]. [2013-12-22]. <http://mooc.net/>.
- [13] The Emory University Libraries. The Emory University Libraries strategic plan. 2012-2015 [EB/OL]. [2013-12-22]. <http://web.library.emory.edu/about/mission-and-strategic-plan>.
- [14] TAYLOR J. Defining e-Science [EB/OL]. [2013-12-12]. <http://www.nesc.ac.uk/nesc/define.html>.
- [15] 雷枫.社会科学的e-Science思考[J].科研信息化技术与应用,2010, 1(2):15-20.
- [16] 师荣华,刘细文.基于数据生命周期的图书馆科学数据服务研究[J].图书情报工作,2011(1):39-42.
- [17] Data Management Task Force. e-IRG Report on Data Management [EB/OL]. [2013-12-20]. <http://www.e-irg.eu/images/stories/publ/>

- task_force_reports/dmtfjointreport.pdf.
- [18] Netherlands e-Science center [EB/OL]. [2013-12-20]. <http://esciencecenter.nl/>.
- [19] 陈春,杨芳,田晓阳,等.科研过程中不同阶段的知识服务需求分析及对策研究[C]//西北五省区第十次图书馆科学讨论会,2010.
- [20] 张红丽,吴新年.e-Science环境下面向用户科研过程的知识服务研究[J].情报资料工作,2009(3):80-84.
- [21] Revolutionizing Science and Engineering Through Cyberinfrastructure [EB/OL]. [2013-12-28]. http://www.astro.caltech.edu/~george/vo/cyber_exec.pdf.
- [22] 马俊才,刘斌,吴林寰,等.利用科研信息化手段推动微生物研究与应用[J].中国科学院院刊,2013(4):519-524.
- [23] 陈孝良,易洪,王小民,等.e-Science(科研信息化)在海洋声学观测领域的应用研究[J].科技管理研究,2010(16):199-202.
- [24] 秦刚,王金一,杨智超,等.e-Science应用中野外数据采集传输系统设计及实现[J].计算机科学,2011(S1):133-135,149.
- [25] 诸云强,孙九林,冯敏,等.论地学科研信息化环境[J].中国科学院院刊,2013(4):501-510.
- [26] KU Libraries Digital Data Services Strategy [EB/OL]. [2014-01-06]. http://kuscholarworks.ku.edu/dspace/bitstream/1808/10603/1/KU%20Libraries%20data%20services%20strategy_01_08_2013.pdf.

作者简介

于倩倩,女,1986年生,硕士,中国科学院文献情报中心助理馆员,研究方向:文献数据库建设。E-mail: yuqianqian@mail.las.ac.cn。

贾茹,女,1989年生,中国科学院文献情报中心在读硕士,研究方向:信息资源组织与建设。

黄金霞,女,1972年生,博士,中国科学院文献情报中心副研究员,研究方向:开放资源建设。

Analysis of Different Research Scene Dependence on Digital Information Resources

YU QianQian¹, JIA Ru^{1,2}, HUANG JinXia¹

(1.Library, Chinese Academy of Sciences, Beijing 100190, China; 2.University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: This paper analyzes the dependence of new academic exchange pattern, e-Science, and Data-intensive Science on digital information resources, and puts forward relevant strategies to handle these new situations. New academic exchange pattern depends on the semantic link, metadata description, fine-grained open, free sharing of digital resources. E-Science depends on metadata description, content storage, links and analysis services and so on. Data-intensive Science depends on massive scientific data, kinds of tools resources and so on. In view of the academic exchanges, we propose to construct diversity information resources especially on open scientific data. To e-Science, we propose to embed open resources into scientific research environment and focus on the different needs of different disciplines. To Data-intensive Science, we propose to enhance the construction of tools resources, and provide more data related services and so on.

Keywords: Academic exchange; E-Science; Infrastructure; Data-intensive Science

(收稿日期: 2014-04-23)