数据密集型科学环境中科技期刊的数字化走向*

李若溪[1)] 游中胜[1)] 田海江[2)] 胡英奎[3)] 黄 颖[1)] 钱建立[4)]
1)重庆师范大学编辑出版中心, 400047;2)重庆邮电大学学报编辑部, 400064;3)重庆大学期刊社,400044:重庆。4)电子设计工程杂志社, 710075,西安

摘 要 数据爆炸促成了科学研究模式的转化,使科学研究过程成为以收集数据—分类处理数据—分析数据为主的"数据密集型"研究。新的科学研究范式对数据的需求是跨学科、跨地域、跨类型的数据大融合,因此各种数据基础设施应运而生。面对这样的改变,科技期刊的应对策略是尽快实现全面数字化、全面开放,推进文献数据的结构化,逐步实现与其他数据基础设施和各种类型科学数据的融合。

关键词 数据密集;科学范式;科技期刊;数字化;开放获取;结构化;融合

The general digitalization trends of scientific and technological journals in data-intensive science environment // LI Ruoxi, YOU Zhongsheng, TIAN Haijiang, HU Yingkui, HUANG Ying, QIAN Jianli

Abstract Data deluge is changing scientific paradigm. Now scientific research is becoming so called "data-intensive science" which more and more based on data capture, curation and analysis chain. Many data infrastructures are emerging as the new scientific paradigm demands integration of multidisciplinary, multiregion and multicategory data. To respond the challenge scientific and technological journals must realize digitalization and open access soon, advance literature more structured and integrate with data infrastructure gradually.

Key words data-intensive; scientific paradigm; scientific and technological journal; digitalization; open access; structured; integration

First-author's address Editorial Department of Journal of Chongqing Normal University, 400047 , Chongqing, China

科技期刊是传播科技信息和知识的,要真正实现这一功能,在数字化、网络化时代,遇到了新的挑战。由于数据激增,促成了科学研究的范式转化为"数据密集型"新范式。所谓"数据密集型科学研究",是指当今的科学研究越来越依赖于数据的聚集和分析,特别是海量数据的分析,在使用计算机的基础上,超大规模的数据工作,已经形成了无处不在的数据环境,我们可把它称为"数据场"(data space)。在数据密集的场

对审稿意见的有效性与可信性进行甄别[16]。编辑部应加强与审稿专家的交流,建立良好的关系,尊重专家的劳动,如及时奉寄与劳动相当的审稿报酬,不要给予太多的审稿任务,每年进行"优秀审稿专家"评选活动,给予表彰和物质奖励等[9]。编辑要做好审稿专家与作者之间的桥梁,共同努力,把稿件修改好。只有审稿专家、编辑、作者三者卓有成效的合作,才能保证论文和期刊的质量。

6 参考文献

[1] 许文深,姚远. 科技期刊审稿的发展[J]. 编辑学报,2001,13(2):70-72
[2] 金晓明. 论学术期刊的审稿方法与发展趋势[J]. 中国科技期刊研究,2007,18(3):372-374
[3] 张惠民. 高校科技期刊审稿方式的优化选择[J]. 宝鸡文理学院学报:自然科学版,2000,20(3):239-241
[4] 孙丽莉,刘祥娥. 高校学报"小同行"审稿专家的遴选[J]. 编辑学报,2011,23(2):139-140
[5] 程静. 选择合适的科技期刊审稿人的几种方法[J]. 中山大学学报论丛,2003,23(5):103-105
[6] 卢圣芳. 对科技期刊审稿专家的选择与搭配[J]. 长江大学学报:社会科学版,2010,33(4):111-112
[7] 刘东信. 综合性科技学术期刊审稿人的选择和外审经验谈[J]. 编辑学报,2010,22(6):521-522
[8] 陈蓉,吕赛英. 科技期刊编辑与审稿专家密切合作的措施[J]. 编辑学报,2005,17(3):203-204
[9] 聂兰英,王刚,金丹,等. 论科技期刊审稿专家队伍的建设[J]. 编辑学报,2008,20(3):241-242
[10] 熊松. 提高科技期刊审稿质量的几点思考[J]. 出版科学,2010,18(4):35-36,8
[11] 曹小春. 学术期刊审稿专家的职责[J]. 编辑之友,2006(6):62-63
[12] 陶范. 审稿专家的责任和权利[J]. 编辑学报,2010,22(6):475-477
[13] 赵丽莹,杨波,张荣丽,等. 对专家审稿的分析和思考[J]. 编辑学报,2010,22(2):146-148
[14] 朱乾坤,石红青. 从审稿统计数据看审稿人的选择[J]. 编辑学报,2010,22(2):151-153
[15] 孙睿老师简历[EB/OL]. [2011-06-13]. http://geog.bnu.edu.cn/teacherweb/sunrui/
[16] 刘荣军. 从信息博弈看科技学术期刊的审稿策略[J]. 编辑学报,2010,22(3):192-194

(2011-07-30 收稿;2011-08-22 修回)

*国家社科基金项目资助(10XTQ007);教育部人文社科基金项目资助(10YJA86001)

中,科研的模式发生着转变,作为科技信息和知识载体的科技期刊,理所当然也会随之而转变,其传统传播方式必须转化,才能适应新的科研环境。

1 “数据密集型”科学研究新范式

与印刷技术的发明一样,计算机技术导致的数据革命正在改变科研的模式。1 000 多年前,科学是靠经验描述自然现象;最近 200 年,科学进入了推理阶段,用模式化方法和综合方法推算、推论自然现象和规律;最近几十年,出现了计算技术模拟自然现象。今天,数据探索式科学,把推理、实验和模拟综合在一起:用仪器抓取数据、模拟产生数据,软件加工数据,计算机存储数据,通过分析处理,得到新的发现[1]。微软研究院的科学家 Jim Gray 把以数据为基础的科学称为科学研究的第四范式[2]。

科学研究正在由假设驱动的科学方法转向基于探索的科学方法。研究者不再设问“我应该设计什么样的实验来验证这个假设”,而是“从这些数据中我能够看到什么相关性”,“如果把其他领域的数据融合进来,能够发现什么新线索”。天文学研究现在不用肉眼去看望远镜,不是对繁星满天或银河闪耀进行直观的分析判断,而是把望远镜观察到的现象以数据形式记录到计算机并对数据进行分析判断。美国的大型天文观察望远镜(LSST)投入运行后第 1 年所生产的数据就达到 1.28 PB[2]。天文学逐渐从一门观察科学变为一门计算科学。除了天文学,其他的科学领域也在发生这样的转变。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%[3]。人类基因组计划 2008 年生产的数据量是每月 1 万亿碱基对,2009 年又增加 1 倍。医学科学的数据激增更是如此,在生物医学文献编目中已经有 1 800 万篇医学文章,现在每年增加近 100 万篇。100 年前,一个内科医生知道医学的全面知识,而今天,一个基层医生需要知道 1 万种疾病、3 000 种药物和 1 100 多种实验室检查才能跟上发展步伐。

数据密集型科学就这样开始了,这是不依个人意志而转移的,是历史发展的必然。面对数据的铺天盖地而来,数据密集型科学研究要做数据抓取、分类处理、分析 3 个基本活动。行使这些功能的数据基础设施和技术也就应运而生了;同时,专门做数据处理、分析的专业人员也应运而生,他们被称为“数据科学家(data scientist)”。

2 数据基础设施与数据科学家

在科研的金字塔中,处于塔尖的大型项目一般有专门的预算用于建立数据和网络基础设施。例如美国的太空计划,其实验预算的 1/2 用于软件费用:一台价值 2 000 万美元的太空望远镜需要几十个人操作它,而它产生的数据却有几千人在写代码。美国和加拿大联合海洋探测计划,费用的 30% 用在网络基础设施上,大约是 1 亿美元。大量的基层科研人员,能够投入研究的软件经费预算可能只能买个 MATLAB 或 EXCEL 等一般的工具。因此,建立通用的“数据基础设施”来为科研服务的需求越来越突出,社会和政府的投入也大幅增加。

2.1 数据基础设施 新的数据基础设施建设涉及到以下关键词:数据分类处理,广泛的无缝链接,数据云,工作流,可视化,等等。

1)数据分类处理。数据录入到计算机中后,首先让程序能够读懂它,也就是说输入的基因、银河系或温度等信息要用算法来重新表述。机器读懂了数据后,对数据分类处理,包含发现正确的数据结构、分门别类、数据转换、图表和元数据长期储存、跨实验跨设施的整合、数据库建模及数据可视化等。目前已有的基础设施项目中,处理数据的能力都达到了 1 TB 级。例如:圣迭戈超级计算机中心(SDSC)建立的数据中心站,拥有 1 PB 以上的数据[4];澳大利亚国家数据服务站(ANDS)的目标是使分散孤立的研究数据转变成相互关联的研究资源。经过了分类处理和整合转换的数据,才能够进行分析利用,才能永久保存和共享,而未经“分类处理”的数据将丢失。

2)广泛无缝链接。在 19—20 世纪,科学数据经常被埋没在科学家分散的笔记本里,或者储存在磁性媒介里,始终未被阅读。这些分散的数据可能随着科学家退休就丢弃了。而今天的趋势是:数据获取、聚集——高效率全天候,跨学科跨国界;数据储存——永久性,动态性,随时读取;数据交流——开放获取,即时互动,世界共享。微软研究院推出的全球望远镜(worldwide telescope WWT)就是个很好的例子[5]。WWT 是宇宙探索工具,聚集了大量星云、星座、行星以及宇宙全景等图像数据,免费提供给用户浏览、研究。通过 WWT, 用户可在桌面上浏览夜空。数据来自哈勃望远镜及分布于世界各地的 10 来个天文望远镜。WWT 处理的数据实现了远程无缝链接:当观察者注意到一个非同寻常的波长或位置的数据,他可以点击那里,而同时远程链接到相关期刊文章上或数据库上;专门为科研人员推出了基于 excel 的数据管理、搜索、转换工具,你可以对自己的 excel 表格中关于天体定位、几何形态等数据直接生成图像[5],也可以链接远程的期刊论文、数据库等等。这些给科研人员节

省了大量重复操作的时间,大大提高工作效率。

3)数据云技术。对付海量数据的管理和加工难题,云计算(cloud computing)是很好的办法。这是一种基于互联网的计算方式,将庞大的计算程序自动分拆成无数个较小的子程序,交由多部服务器组成的系统进行搜索和计算,最后又将处理结果返回给用户。能够在数秒时间内处理数以亿计的信息。例如最近出现的微软卫生库(Microsoft HealthVault)[6]和谷歌卫生(Google Health)[7],都是基于因特网的"用户数据云",临床病人的数据输入里面形成云。用户数据云为新医学知识即刻传达至病人提供了可能。维基百科也是类似的用户数据云。

4)工作流技术。这是对工作流程及其各操作步骤业务规则的抽象、概括、描述。工作流要解决的主要问题是:为实现业务目标,在多个参与者之间,按预定规则自动传递文档、信息或任务。它的好处是有利于管理数据,对纷繁复杂的数据处理和分析起到提高效率减少差错等作用。

2.2　数据科学家　美国国家科学委员会(national science board NSB)在"长期保存数字化数据集成:推进21 世纪的研究和教育"计划中,提到对"数据科学家"这一新群体的关注和扶持问题[8]。所谓的数据科学家,包括信息与计算机科学家、数据库和软件工程师、学科专家、数据处理员和专业注释员、图书馆员、档案馆员等从事数据集成的管理人员,科技期刊的编辑人员当然也在此列。在未来的科学研究领域内,多了一个特殊的群体——数据科学家,计算机专业的人将大显身手,而缺乏计算机基础的人在这里无立足之地。期刊的编辑将面临这样的挑战。

3　在数据密集型科学环境中科技期刊的应对策略

期刊文献也是数据资源类型之一。现时的文献信息往往是分散地孤立存在,或者说处于非可融合状态。如果期刊文章继续保持分散状态,其传播功能、储存功能将逐渐萎缩,而评价功能也将在数据密集科研的盛行中失去意义。

这一点已经从目前的发展苗头看出,因此,期刊的出路在于与时俱进,全面数字化,全面开放,推行结构化,推进文献与数据的融合。

3.1　全面数字化和全面开放　早在 2003 年,数字化信息就占据了人类有史以来所创造的全部信息的90%,超过了历史上纸质和胶片信息的累计量总和[2]。目前我国的科技期刊,绝大多数都做到了分散数字化出版,也就是把电子文本提交给 CNKI、万方、维普等大型数据库,部分期刊已建立自己的网站。据调查[9],中国 1 800 多种科技核心期刊中,1 100 多种有自建的网站,占 59%。本课题组的调查结果显示:中国大陆学术期刊(即中国知网的 8 100 多种来源期刊),有自建网站的占 49%;国外学术期刊有自建网站的比例为 73%。这是很好的数字化措施,但还不能算是全面数字化。

2001 年以来国际上兴起的开放获取(OA)运动,在很大程度上促进了期刊的数字化和开放。现在美国所有的公共资助的生物医学科学文献必须在线开放于 PubMed Central 中心知识库[10]。欧洲发达国家也纷纷跟进。瑞典 LUND 大学的 OA 期刊目录 DOAJ 收录的期刊数已超过 7 300 种[11]。

虽然我们把研究成果出版,但出版的文献仅仅是全部研究数据的冰山一角。期刊数字化,是要达到文献与所有科学数据能够相互融为一体,在英特网上形成数据与文献互动操作的世界平台,这才算是全面数字化。

3.2　数据与文献的融合　文献[2]中阐述了科学数据的整体组成,类似于金字塔形:文献数据处于塔尖;基础层是大量的原始数据;中间层是抽取出来的和关联的数据层。这 3 个部分在数据场中相互融合,共同有机地构成了全部科学研究的内容整体;如果 3 部分相互独立,就如同有机体变成断肢残臂,大大削弱数据资源的价值。

所谓融合,就要达到:在构建的数据平台上你可以读一篇论文,而同时调取它的原始数据;你甚至可以重演作者的分析过程;或者你能够在分析一些数据的同时找出跟数据相关的全部文献。Entrez, 是一个生命科学搜索引擎,真正实现了数据和文献的交互性操作,用户可以边阅读一篇文章,同时打开基因数据,跟随基因找到这个疾病,然后又回到文章[2]。微软的 WWT,也实行了数据与文献的融合。

融合和交互操作发展了一些统一的链接、统一的标签和 ID 号,技术上能够实现把全世界的数据都集成在一起,形成巨型的动态数据集,一个全球化的数据库将必然诞生。

3.3　文献内容结构化　期刊文本型信息是半结构化的,数据库的数据是全结构化的,要把二者融合到一起,数字化、开放是必要的前提,推进文本的结构化是必要的手段。对文献文本的结构化标准化处理,目前已有 2 方面的措施。

1)自动化标引。期刊文献标引工作如果得到推广和统一,那么文献信息的结构化就必然增强,与结构化的数据库之间的交互性操作将会更为深入和流畅。

标引工作在计算机数据处理中属于语义服务,由语义服务指导数据工作者提炼数据,利用自动工具在文本和数据库中形成语义层通道,从中可望为数据的处理分析和整合提供很有效的解决途径。英国皇家化学学会《分子生物系统》杂志(Royal Society of Chemistry's journal Molecular BioSystems)[2],对HTML格式的全文内有关化学物质叙词进行标注,并把这些标注的词汇链接到外部数据库词目。标引工作由具有叙词标引专长的编辑来完成,借助自动化文本挖掘工具(一种能从文本中提取词语的工具软件)的协助。出版环节的标引恰恰是出版增值服务的体现。

2)先进的文本分析技术。期刊信息一旦融合到数据场中,将帮助科研人员解决单靠阅读文献无法紧跟学术动态的问题。先进的文本分析技术,侧重于提高文本的机器易读性,例如用文本分析技术从文献中抽取实体(entity)和实体之间的关系(entity relation),利用机器定义和识别的语词,嵌入文献中,使过去只能供人工阅读的文献就能够用机器来分析,让机器去寻找不同学科的文献之间的关联点,从而串联知识点,触发新视野的产生,获得无法预想的跨学科推理。

美国的一些研究项目鼓励学者们在出版论文时就发布实体或实体关系信息,以尽量减少后加工过程[2]。

科技期刊的文本内容,本来就是半结构化,例如摘要4要素、量和单位、参考文献等,但要达到机器可识别的文本分析,结构化工作还有很长的路要走,正如我们现在的"标准化"一样,虽然几乎付出了整整一代编辑人的努力,但仍然还有很多非标准化的死角。

3.4 基于网络和数据场的学术过程记忆 传统的文献中的引文,可以引导读者找到相关的更广泛的信息;它记录了学术进步的印迹。同时,引文系统主导着论文和期刊的学术评价。在数据密集型科研环境下,引文索引和评价将不再起主导作用。因为数据场中信息的类型、来源渠道和获取方式都是多元的,各种数据的流动、交互操作、融合、引用等等,都将留下轨迹,因而在网络中记载和显现这种过程就成为可能。有关专家正尝试使学术过程以机读信息发布于英特网,称为"过程公开记忆",把隐性的数据流动转变为显性的,甚至可视化。

例如,出版物的引文体系可以建立起这个过程的路径图像。基于网络和数据场的学术过程,与传统引证体系相比,更加立体化、综合化,将在学术跟踪和评价中大显身手。

4 结束语

数据密集型科学范式正在出现。让所有的科学文献都在线,所有的科学数据都在线,它们之间能够交互操作,已经成为时代的呼声。当大量对科学的引擎起燃料作用的数据游离于期刊论文之外的时候,我们再用大量的人力物力去加工论文文本,而忽略那些需要花力气去处理的数据,我们可能又将落后于西方发达国家了。因此,我国的科技期刊要尽快做到全面数字化,推进结构化,主动与数据基础设施融合,最大限度地实行OA,才能顺应潮流,应对挑战。

5 参考文献

[1] Gray J. eScience: A Transformed Scientific Method [EB/OL]. (2011-08-20)[2011-09-07]. http://research.microsoft.com/en-us/um/people/gray/JimGrayTalks.htm

[2] Hey T, Tansley S, Tolle K. The fourth paradigm: data-intensive scientific discovery[M]. USA,[出版地不详]:Microsoft Research, 2009

[3] EMBL-EBI. EMBL-bank Growth[EB/OL].[2011-09-07]. http://www.ebi.ac.uk/ena/about/statistics

[4] San Diego Supercomputer Center. Cyberinfrastructure[EB/OL]. [2011-09-17] http://www.sdsc.edu/about/Infrastructure.html

[5] Microsoft Research. What is WWT? [EB/OL]. (2011-08-09) [2011-09-17]. http://www.worldwidetelescope.org/whatIs/whatIsWWT.aspx

[6] Microsoft. Microsoft Health Vault[EB/OL]. [2011-08-30). http://www.microsoft.com/en-us/healthvault

[7] Google. Google Health[EB/OL]. [2011-08-30). http://www.google.com/intl/en_us/health/faq.html

[8] National Science Board. Long-lived digital data collections: enabling research and education in the 21th century[R]. USA: NSB-05-40, 2005

[9] 程维红,任胜利,路文如,等.中国科技核心期刊网站建设现状[J].中国科技期刊研究,2011,22(5):649-655

[10] U.S. Department of Health and Human Services. National Institutes of Health Public Access Policy Detail[EB/OL]. [2011-08-30]. http://publicaccess.nih.gov/policy.htm

[11] Lund University Libraries. Directory of Open Access Journals as of Today[EB/OL].[2011-09-18]. http://www.doaj.org/

(2011-09-20 收稿;2011-10-20 修回)