大数据视域下的科技期刊数据库建设*

刘 俊 张 昕
清华大学出版社期刊中心,100084,北京

摘 要 在总结大数据特点的基础上,梳理各科技期刊数据库已经进行的数据收集、数据整合、数据利用等具有前瞻性的基于数据的实践。认为今后科技期刊数据库应注重数据的精化和标准化,积极向信息服务商转型,并始终以大数据的思维来指导实践,从而更好地顺应大数据趋势。

关键词 大数据;数据收集;数据整合;数据利用;数据精化和标准化;信息服务商

Construction of scientific journal databases from the perspective of "mega-data" // LIU Jun, ZHANG Xin

Abstract The characteristics of mega-data are summarized by sorting the forward-looking practices of foreign scientific journal databases in data collection, data integration and use of data. It is pointed out that, in the future, scientific journal databases should attach importance to refinement and standardization of data, change the role of a content provider to an information service provider positively, and take the idea of mega-data as the guide of practice, in order to adapt to the trends.

Keywords mega-data; data collection; data integration; use of data; refinement and standardization of data; information service provider

Authors' address Journal Publishing Center of Tsinghua University Press, 100084, Beijing, China

如果说过去人们只能通过数据总结历史规律,凭借既往经验作出新的决策,那么从今以后人们则可以通过数据预测未来,拥有"先见之明"来把握先机、规避风险;而实现这种从事后到事前、从被动到主动的质的转变的正是大数据。

1 大数据之"大"

数据之于我们并不陌生,它渗透于我们学习、工作、生活的方方面面,但此数据并不等同于大数据中的"数据"。何谓大数据? 它是指"大小超出了传统数据库软件工具的抓取、存储、管理和分析能力的数据群",体现为数据体量大(Volume)、数据类型多(Variety)、处理速度快(Velocity)、价值密度低(Value),即4V特征[1]:1)数据体量大。指数据的量逐层递增,没有穷尽。2)数据类型多。包括结构化数据和图片、音频、视频、动漫等非结构化数据。3)处理速度快。指以并行工作方式来高效、精准地处理海量数据。4)价值密度低。指真正具有利用价值的只是数据中很小的一部分。

2 科技期刊数据库基于数据的实践顺应了大数据的时代浪潮

大数据的4V特征更加彰显了数据的重要价值,大数据时代的到来引发了世界范围内的广泛关注。大数据技术为各行各业带来了新的发展契机。如今,在商业、金融、医疗、科研、教育、政治等领域,行业先行者们正借力大数据技术创造着新的更大的价值,而科技期刊作为最前沿科技动态的传播者,也必然是最先对大数据予以关注的行业之一,其中,科技期刊数据库是核心力量。当然,这些实践并不是在大数据概念提出之后才产生的;因为虽然大数据的概念是新的,但是基于数据的实践却早已有之,大数据时代的到来恰好印证了这些实践的前瞻性与创新性,我们不妨以大数据的视角来审视这些实践。

科技期刊数据库首先需要扩大数据总量、丰富数据类型,这是利用数据创造价值的前提和基础;在此基础上,科技期刊数据库需要整合数据、优化数据,这是充分利用数据、使数据价值最大化的保证;最后通过分析数据有的放矢地利用数据,这是最根本的目的,也是大数据的真正价值所在。这3个阶段逐步深化,不可逾越,唯其如此,科技期刊数据库才能更加理性、科学地应对大数据浪潮。国外科技期刊数据库在这些方面所进行的探索可圈可点。

2.1 收集数据——扩大数据规模 收集数据是最基础性的工作,以此扩大数据规模,包括增加数据容量和丰富数据类型2个方面。

2.1.1 *链接其他文献数据库* 科技期刊数据库在不断充实自有资源的同时,可以链接其他文献数据库,以达到扩大数据规模、共享数据资源的目的。例如,施普林格平台利用DOI技术实现了与300多家出版商的期刊和专著的链接,用户通过PubMed、ChemPort、CrossRef等链接可以获取参考文献书目文摘乃至全文[2]。又如,爱思唯尔旗下的SciVerse平台不仅整合了ScienceDirect、Scopus和相关科技网页上大家所熟知的、备受信赖的高品质内容,还具有前瞻性地添加了第三方开放的创新性工具和应用程序,从而丰富、扩展了原有内容的价值[3]。

* 教育部科技发展中专项研究资助课题(FSSP 2012110)

2.1.2 *建立过刊数据库* 国外科技期刊数据库不但及时将新资源补充进来,还非常重视过刊资源,建立过刊数据库,从而使数据资源具有时间的连贯性,更加完整,易于查找。例如,施普林格从2004年开始将所有过刊进行数字化转换,到2008年,施普林格回溯期刊库包括11个学科的超过3万期的920余种期刊,共3万多期,由11个重要学科包组成,其中绝大部分期刊从第1卷第1期开始提供,有些期刊的访问年限远至1854年[4]。又如,爱思唯尔自2001年1月启动回溯文档项目,1995年以前全部期刊的文章均可进行全文检索,并可显示HTLM格式的文摘和参考文献列表,还能直接链接至被引文献[5]。

2.1.3 *丰富数据类型* 大数据不仅是数据量的爆炸性增长,还包括数据类型的变化。过去的数据都是以二维形式存储于数据库中的结构化数据,如期刊数据库中存储的文本文件、Excel表格处理的数据等。随着多媒体技术的发展,有些期刊数据库逐步丰富数据资源的类型,增加图片、音频、视频等非结构化数据。例如:《Nature》自己开发各种科学奖颁奖仪式的在线视频;《Nature Chemical Biology》采用3D技术将分子结构用三维方式展现出来[6];《柳叶刀》的平台上每期都有一个不超过20 min的专题视频介绍,以及简要介绍当期杂志内容的音频(MP3)[7];SpringerLink实现了书刊互连支持流媒体技术,提供能够缩放的高清晰图片、动态表格、三维动画[8]。此外,还有视频期刊——《可视化实验》,它是一种以视频为主要媒介(而非补充材料)的在线期刊,其文章拥有一个简短的视频片段,只管记录实验所需步骤,这一期刊已经引起生命科学领域的一定关注[9]。

2.2 整合数据——提高数据质量 数据收集工作完成后的整合工作,也是必不可少的,否则,混乱、零散的数据将使其利用效果大打折扣。不少科技期刊数据库都在进行数据格式化、标准化的工作。

2.2.1 *数据标引* 从微观层面来说,将文献数据碎片化后再标引,是期刊数据库内部进行数据格式化的主要方式。目前,业界普遍认为XML语言适合科技论文的标引,这是因为XML能够实现"一次排版,多媒体发布",XML标引是网络出版最基本的技术支撑。此外,《Mendeley》建立了一个结构化的数据库,负责提取数据和全文以让这些文件能被全文注释和标注,这些数据也能被研究团队分享和讨论[10];爱思唯尔未来的文章都将带有标签格式,能够使读者在文章中准确定位[10];洛克菲勒大学出版社重视图片元数据的标引和呈现,从2002年9月开始,用Photoshop对作者的图片逐一进行扫描分析,一旦发现图片造假,文章将不被录用[11]。

2.2.2 *数据库平台的无障碍对接* 从宏观层面来说,各期刊数据库之间格式兼容、无障碍对接,能够实现资源最大化,有利于资源的进一步深度开发和利用。例如,CrossRef主要利用DOI(学术与专业出版物数字对象标志系统)来实现不同出版社在线学术资源之间高效而可靠的交叉链接,通过CrossRef,图书馆无须跟各个出版商签订双边协议,也无须跟踪出版商的链接表,就能在CrossRef成员出版商之间建立链接[12];EndNote是SCI的官方软件,支持3 776种参考文献格式、几百种写作模板,涵盖各个领域的杂志,能直接连接上千个数据库,并提供通用的检索方式,其他国家的期刊数据库下载数据时,均支持EndNote[13];HighWire Press期刊被PMC、Scirus和Google等多个仓储和搜索引擎收录、检索与链接,用户可通过PMC数据库链接获取HighWire Press平台发表的所有研究论文[14]。

2.3 利用数据——体现数据价值 数据收集和整合都是基础性工作,若想使数据发挥更大作用,就要进行数据分析,有的放矢,提取有价值的资源,摒弃垃圾资源,这是大数据真正价值所在。对于数据的利用主要是通过受众行为数据分析得出的成果来优化用户服务。

2.3.1 *调整期刊内容* 期刊数据库可以根据用户的访问行为数据来了解用户的喜好,据此调整期刊的内容。例如,国外医学期刊统计网络点击、访问、页面浏览、文摘浏览、目录浏览等数据建立用户数据库,通过这些数据归纳用户最感兴趣的主题、搜索最多的词,并根据用户的不同国籍、不同年龄、不同专业进行分层分析,以此来动态调整期刊的报道重点[15]。

2.3.2 *建立学术互动机制* 论文作者可以借助期刊数据库提供的软件工具来获得关于用户基本情况的数据,以作出更及时、更有针对性的反馈,形成良性互动。例如,著名的学术研究网站Academia. edu发布了一款实时追踪分析工具,科研工作者借助这款工具可以进行具体的学术受众分析,从而了解自己的研究方向受到哪些地区、哪些人群的关注,从而可以进行积极的学术互动和研究反馈,形成良好的学术交流互动的网络氛围[16]。

2.3.3 *改变学术评价体系* 受众行为数据分析可能改变期刊界的学术评价体系。过去,学术论文录用与否一般由编辑或审稿专家来决定,评判的指标是定性的、主观的;而受众行为数据分析的引入,将可能使读者的范围和数量也成为重要的评判指标,这是定量的、客观的评判。例如,Limerick大学的Tim Ritchie教授就对此表示:"在我申请升职或说明自己的学术研究成果时,我会把我的网络读者受众分析报告打印出来,我想这会成为一个重要的评判指标。并且这些数据会鼓励我进行学术研究,我也能很好地知道自己学术研究的受众是

哪些,哪些学术研究是被大众所认可的。”[16]

2.3.4 语义查找便于用户准确定位信息　语义查找技术就是利用关键词的内在逻辑联系,通过权重计算来提供用户最需要的信息。例如,爱思唯尔的语义技术和知识发现软件 Collexis 的语义查找技术功能强大,用户在面对海量数据时可以很便捷地找到自己的同行及其在世界各地的分布情况以及联系方式,极大地促进了研究者之间的研究协作和信息资源的交流,还可通过关键概念进行搜索并提供科学内容,基于个人获得的资助项目或出版物显示出相关的学者、资助和出版物[17]。

3　大数据环境中科技期刊数据库应更加主动地顺应趋势

上述实践并非完全产生于大数据成为热潮之后,很多都是期刊数据库跟踪技术发展趋势、满足用户需求的顺势之举,而这些实践正顺应了如今的大数据浪潮。这也从一个方面说明,大数据时代的到来有着扎实的现实基础,是技术发展的必然趋势。当大数据的概念逐渐明确、大数据时代的轮廓逐渐清晰之后,科技期刊数据库就应该更加自发、自觉地调整决策,顺势而为。

3.1　数据在多更在精　大数据的首要特征是数据量大,这在一定程度上使期刊数据库热衷于“跑马圈地”,不加选择、没有区分地将各种资源纳入囊中;同时,期刊出版、用户体验、数据库营销等各种行为产生的海量数据必然包括大量无用的垃圾数据。这些从主观和客观上都造成了数据存储和处理的危机,也必然增加了成本;因此,数据容量大也是一把双刃剑,应该理性、客观地看待大数据的大容量。目前,期刊数据库更加关注掌握了多少期刊资源,而缺乏对已有数据分析利用能力的重视,这就陷入了误区。由于数据分析才是大数据的真正价值所在,所以,面对大数据,要“去粗取精,去伪存真,取其精华,去其糟粕”,这样才能保证大数据真正发挥作用。

3.2　自上而下与自下而上的力量相结合实现数据标准化　要真正实现大数据潜在的大价值,数据标准化是基础。目前,期刊数据库的数据格式混乱、期刊数据库直接对接的问题还比较突出,这必然会降低数据分析的效率,影响其价值的发挥;因此,需要一股自上而下的力量,以政府机构、行业组织、大型期刊出版机构为主导,以公共力量推动数据的标准化、客观化、透明化,可以制订统一的数据标引标准,自上而下贯彻。另一方面,通过自下而上的力量,使数据的碎片化、格式化和数据标引真正落实,比如通过“众包”这一具有鲜明大数据时代特色的生产组织形式,将数据标引这种细琐而质量要求高的业务外包给机构之外的小型组织或个人,通过最基层的力量完成最基础的工作,再由上层的机构进一步将数据予以汇总和整合。

3.3　加快从期刊内容提供商向期刊信息服务商的转型　大数据的价值体现在 2 个方面,分析使用和二次开发;因此,如果期刊数据库仅仅满足于不断扩张自身的规模,那么其始终会停留在期刊内容提供商的层面,即使采取了一些开发新终端、增加科技资讯的举措,也还是仅仅停留在信息资源服务的浅层次。在大数据时代,科技期刊数据库要做的是充分挖掘数据,将学科的发展脉络、学科之间的关联梳理清晰,将与学科相关的科研机构、科研人员信息梳理清晰,将数据库运营、营销等各个环节的信息梳理清晰,这是并行的几张网络,在此基础上建立网络的纵向关联,由此形成一个多维的数据结构。此时,科技期刊数据库不但可以追踪最新的科研进展,还能为用户提供具有针对性的个性化服务,并能为政府等相关部门的决策提供信息参考和技术支持。如此,科技期刊数据库才可以说实现了对于数据的深度挖掘和利用,实现了期刊信息服务商的真正转型。

3.4　始终以大数据意识来开展各项实践活动　大数据之“大”终究是一个相对概念,随着科学技术的突飞猛进,今日之“大”也许会成为明日之“小”;因此,重要的是要具有大数据的意识,大数据并不仅仅是大型期刊数据库平台的专利,小型数据库甚至单刊都可以在大数据时代有所作为。例如创办于 2012 年 7 月 12 日的大数据期刊——《GigaScience》,采用标准全文文献、数据库信息以及信息分析工具相结合的崭新模式来发表大规模的生物学研究成果,读者不仅可以获得文献中所得到的科学结论,还可直接通过文献所提供的数据和分析工具对结果进行测试和验证,实现了数据的透明、公开及可重现性,这是传统期刊出版业迈向数据全面公开与共享的重要一步;此外,其数据库 GigaDB 已采用数字对象唯一标志符(DOI)对期刊数据库中的所有数据进行标志,使数据保存更加永久,实现可追踪、可检索、可链接、可引用,而之前这些功能仅能用于参考文献[18]。该例中,单刊也进行了大数据的实践,一是实现数据的透明公开,二是实现全部数据标引,走的是一条集约型的大数据之路。由此可见,即使单刊也可以进行基于大数据的实践,数据量只是一个方面,更重要的是拥有大数据的眼光和思路。

总之,大数据浪潮来袭,科技期刊数据库应以积极、理性的态度应对这次科技浪潮,实现扩大数据规模与提高数据质量的平衡,扩展外延,夯实内涵,实现数据的深度挖掘和充分利用,这是大数据的题中之义。

4　参考文献

[1]　中国电子报,电子信息产业网. 大数据的四个典型特征

[EB/OL]. [2013-10-25]. http://cyyw.cena.com.cn/a/2012-12-04/135458292978407.shtml
[2] 丁岭. 施普林格数字出版发展模式探析[J]. 大学出版, 2008(2):62
[3] Elsevier. 电子产品信息:SciVerse[EB/OL]. [2013-10-25]. http://china.elsevier.com/ElsevierDNN/%E7%94%B5%E5%AD%90%E4%BA%A7%E5%93%81%E4%BF%A1%E6%81%AF/SciVerse/tabid/1594/Default.aspx
[4] 湖南师范大学图书馆. Springer 在线回溯数据库(OAC)简介[EB/OL]. [2013-10-25]. http://lib.hunnu.edu.cn/szzy/ShowArticle.asp?ArticleID=7
[5] Elsevier. ScienceDirect 期刊回溯文档[EB/OL]. [2013-10-25]. http://taiwan.elsevier.com/elsevierdnn/sm/tabid/1066/default.aspx
[6] 张聪,张文红. NPG 期刊运营特点分析[J]. 科技与出版, 2013(2):8
[7] 任胜利. 柳叶刀(Lancet):医学期刊中的锋利之刀[EB/OL]. [2013-10-25]. http://blog.sciencenet.cn/home.php?mod=space&uid=38899&do=blog&id=247729
[8] 陈丹,程小雨,齐媛媛. 施普林格期刊运营模式及数字出版策略分析[J]. 科技与出版,2013(2):17
[9] 梅尔沃德. 视频期刊:科学界的 Youtube[EB/OL]. [2013-10-25]. http://www.bnump.com/news.php?id=10000684
[10] 甘兹特. 技术创新为 STM 出版以及科研提供了新工具[EB/OL]. 郑珍宇,译. [2013-10-25]. http://www.bookdao.com/article/40068/
[11] 张学东,赵爱群,杨雷,等. 美国洛克菲勒大学出版社考察纪实:兼对我国科技期刊办刊体制的思考[J]. 内蒙古民族大学学报:社会科学版,2011,34(4):123
[12] 周庆辉,陈红云,张晶. DOI 与 CrossRef 在科技期刊出版中的作用[J]. 编辑学报,2009,21(1):68
[13] 百度百科. EndNote[EB/OL]. [2013-10-25]. http://baike.baidu.com/view/959995.htm
[14] 史海娜. HighWire Press 期刊平台研究[J]. 出版科学, 2009,17(3):90
[15] 马英,胡永成. 谈国外医学期刊经营中的服务理念[J]. 编辑学报,2011,23(6):560
[16] 韦龑. 网络正在改变学术界[EB/OL]. [2013-10-25]. http://www.bookdao.com/article/44978/?type=98
[17] Elsevier 发布针对学术用户的软件 Collexis:将 16 个机构的研究人员联系起来[J]. 现代图书情报技术, 2010(11):44
[18] 大数据期刊《GigaScience》喜迎创刊一周年[EB/OL]. [2013-10-25]. http://www.genomics.cn/news/show_news?nid=99611

(2013-11-01 收稿;2013-11-20 修回)

复句内部不应当用句号

复句有 2 类:一类是简单复句,即复句中只有 2 个单句;另一类是多重复句,即复句中有 2 个以上单句。

GB/T 15834—2011《标点符号用法》4.6.3.2 规定:“表示非并列关系的多重复句中第一层分句(主要是选择、转折等关系)之间的停顿”,应用分号。可见,多重复句内部用句号是不规范的。为便于理解,标准列举了如下 4 个示例:

示例 1 人还没看见,已经先听见歌声了;或者人已经转过山头望不见了,歌声还余音袅袅。

示例 2 尽管人民革命的力量在开始时总是弱小的,所以总是受压的;但是由于革命的力量代表历史发展的方向,因此本质上又是不可战胜的。

示例 3 不管一个人如何伟大,也总是生活在一定的环境和条件下;因此,个人的见解总难免带有某种局限性。

示例 4 昨天夜里下了一场雨,以为可以凉快些;谁知没有凉快下来,反而更热了。

我们在加工《编辑学报》的稿件时,发现不少作者在多重复句内部用了句号(也有在简单复句内部用了句号的,但不多见;简单复句中,分句之间的停顿一般用逗号),甚至我们订正后,把加工稿发给作者校核认可,大多数人在返回的校样中,把我们订正过的又改回去。这说明这些同人没有掌握标准规定的用法。例如:

1)如果不认真把关,就会造成重大的政治性错误,造成难以挽回的损失;因此,高校科技期刊办刊人员必须保持清醒的头脑,要有高度的政治觉悟,坚守政治思想这一阵地,时时刻刻都不能掉以轻心。

2)应该看到,每篇稿件都承载着作者的心血与期待,不论稿件的学术水平、应用价值、技术含量怎样,都是作者科研劳作的直接体现;所以,科技期刊应对事业、对作者、对读者负责,秉持严谨负责与公平公正的态度,认真对待每一篇稿件。

3)OA 运动旨在促进学术成果的共享和自由交换,构建一个真正服务于科学研究的学术交流系统;然而,学术期刊属于准公共产品而非纯公共产品,其出版活动中除了存在社会公共需求,还存在着作者、出版者各方的私人需求。

4)虽然我国科技期刊大都有明确的投稿要求和严格的编辑规范;但是,由于长期以来在初审时对技术审查工件重视不够,导致部分作者投稿时漠视所投刊物的要求,稿件格式极不规范,其随意的投稿态度与向国外投稿的谨慎态度相去甚远,因此,初审中强调技术审查,能使作者端正投稿态度,更加重视文章的规范表达。

这 4 个句子中,分号用得对,而改成句号就错了。

(同 任)