专业文献数据库的质量控制

江 洪

(中国科学院武汉文献情报中心 武汉 430071)

摘 要 专业文献数据库开发研制中的质量控制包括:数据采集、加工质量、检索软件的质量、产品载体形态等方面。本文阐述了专业文献数据库建立中,质量控制的一些体会。

关键词 专业文献数据库 质量控制

数据库建设是信息社会的一个重要组成部分,是方便有效地为社会提供信息的重要渠道。一个数据库就是一个完整的检索系统,它的质量包括:数据质量、检索软件的质量、产品载体形态等。其中数据质量是数据库质量的核心,它直接影响数据库的检索性能。而影响数据质量的因素有:信息源采集的质量、数据规范及数据加工的质量。在编制"长江流域资源环境科学文献数据库"的实践中,我们对提高专业文献数据库质量有了一定的认识。

1 与文献单位的特色服务建设和专业期刊结合,突出专业特色,提高数据的质量

专业文献数据库的特点在于它的数据专业性,所采集的数据比较集中地反应出该专业领域研究的全貌。专业人员通过它的检索可比较方便和全面地掌握该研究领域的文献线索,从而对他们的工作提供有益的帮助。我们所建的"长江流域资源环境科学文献数据库"的专业范畴涉及到各类资源(水、土地、生物、气候、能源、人力、旅游等)的开发利用中的科学问题,以及保护资源、开发可更新资源的途径、环境变化对生物多样性的影响、生态保护与建设、环境保护与治理、自然保护区、各类自然灾害、大型工程的环境影响与建设、沿江产业带建设、城市与城市化、农业农村发展、交通运输网建设、区域社会经济战略等。由于长江流域的开发是90年代中国经济发展的重大战略部置,这些资源和环境问题近些年来已成为全国乃至全球关注的热点。

从所涉及的范围来看,其研究是综合性的、交叉的。为了确保数据的专业质量,准确确定"长江流域资源环境科学文献数据库"的数据收录范围,我们首先与中国科学院和中科院武汉文献情报中心的研究方向、服务特色相结合。中科院一直非常重视生态建设以及协调人类与自然环境关系的研究,1995年编制了《中国21世纪议程中国科学院优选项目计划》,作为院"九五"科研计划的重要组成部分,同时提出了"突出区域与学科特色,发展综合优势"的措施,将环境保护及资源开发、生态建设与社会可持续发展进行系统综合研究,已有一大批研究所在进行长江流域生态与环境方面多学科的综合研究,大大促进以三峡和长江产业带开发带动的"长江流域生态环境研究基地"的形成与发展。宏观环境使我们的数据库有了发展的良好条件。同时,通过多年的情报调研,创办《长江流域资源与环境》学术刊物;调整馆藏结构,中科院武汉文献情报中心已初步成为具有长江流域资源环境文献情报服务特色的文献情报中心。我们以本中心的文献收藏为依托,确保了数据的完整性和专业质量。而作为一项重要的基础性文献情报工作,"长江流域资源环境科学文献数据库"的研制将为本中心形成学科特色方面增加一个重要的内容。

由于本课题组成员包含了《长江流域资源与环境》杂志编辑部的全体成员,在确定范围时,我们结合杂志的报道内容和报道范围,并参考《资源科学主题词表》的涵盖范围,对采集的数据由该刊编辑人员进行专业质量控制。还通过广泛征求杂志编委、读者、作者的意见,及时补充数据,完善数据库,把握其专业质量。

2 传统建库方式与套录建库方式相结合,加强数据加工、规范质量的控制

建设一个数据库,要投入大量的人力、物力和财力。数据库的建设者们都是经历了从数据采集、标引等各种艰辛劳动而最终建成库的。我们在建库工作中经常可以发现我们所建之库与前人已建成之库存在或多或少的重复,这种情况在这种专业文献库中尤其常见。通常专业文献数据库所需各类资料大多已蕴含于各类大型综合文献库中,这样重复就是巨大的,对于这部分重复的数据我们没有必要也不应该再花费等量的艰辛重做一遍,而采用套录这种手段可使这部分信息资源得以极大地共享,节约资金,减少重复

劳动。因而套录成为我们首选的建库手段，不仅克服了资金少、技术力量薄弱的缺点，还使现有信息资源得到充分的共享。

套录不是万能的，它毕竟只是一种手段，而且有其适用范围。套下的文本数据既不规范，也没有检索功能，必须经过软件人员的后处理，配以相应的检索软件，才能使之得到真正的使用。所以套录并不意味着建库工作的全部或大部，仅仅是建库的开始。

我们采用了各种检索途径和手段，对现有的相关商业化公开出版的数据库进行检索和跟踪，选择套录的数据库有8个，它们都是正式出版的电子出版物，其数据质量都经过各自严格的控制。由于各个数据库有其各自的特点，在数据格式和标引深度等方面是很不相同的。即便是套录下来的文本文件，格式也是千差万别，要对套录下来的大量数据逐条进行检查、格式转换、归并、去重、增减字段、编辑等等处理，使数据获得统一标准的格式，产生新的组合，建立新的结构，才能入库。这与一般意义上的套录有着很大的区别。

除套录建库以外，我们还利用本中心的馆藏文献，进行手工补充数据，同时根据有关标准，对专著、会议录等进行逐条手工制卡和录入。

在采集数据加工过程中，我们强调质量控制，有人负责专著的手工检索和录入；有人负责新现刊的跟踪和手工标引录入；有人负责现有数据库的检索与套录，对套录的数据进行复杂的格式转换和规范处理。再经过去重，然后将数据随时交给审核人员进行质量审查，审查后的合格数据交负责数据库的维护人员入库、调试和试运行，进行最后的修正、数据合并和备份处理。每一条数据都要经过这个流程，使得数据质量得到了有效的控制。

3　编制高质量的检索软件

检索软件的质量要求数据库占用空间省，对用户机型及汉字系统兼容性好，安装简捷、快速，检索速度快，方式全，用户界面友好，检索辅助功能多等。数据库研制课题组中有多年从事计算机检索工作的人员，他们在工作中使用和接触了国内外大量的文献检索系统软件，对各个系统的功能和优劣有深刻的认识。在编制开始阶段，首先选用那些投资少、水平高、功能强、使用方便、对硬件系统要求低的商用管理软件。通过仔细的调查，反复的比较，最后我们选定了中国人民解放军医学图书馆开发的“中英文参考文献管理员”来作为我们的系统软件。

随后，我们又继续集中力量，研制开发自主版权的高水平数据库管理系统软件，由原来的DOS版本系统软件改为WINDOWS版本系统，屏幕更美观，操作更直接，界面更友好。该系统对运行环境的要求不高，可以在486DX66以上各档次微机及其兼容机上运行，软件环境为WINDOWS操作系统。系统的主要特色有：多功能、多途径、一体化、标准化、科学性、完整性、实用性、通用性。系统主要有全文检索、概念组配检索功能、记录可选择打印、存盘输出功能、一次存盘无最高数量限制、WINDOWS界面、检索进程指示等特点，功能上向通行的检索系统看齐，操作上与热门的应用软件类似。本系统具有的关键词检索功能非常实用，无需使用截词符，可方便的在“ti”、“ab”、“kw”字段进行查找。

4　开发多样化的数据库产品

由于我国信息网络建设还处在起步阶段，数据库的联机网络服务在数据库服务中所占比例很少，而大量使用的是以分散模式存在的脱机数据库服务方式，这要求数据库产品必须适应市场需求，实现多样化。我们已推出的产品主要是软磁盘型，并将尽快推出CD—ROM光盘型、网络版。

软磁盘型数据库以软磁盘为数据库存储介质，具有数据更新快，成本低的特点，非常适合较小的专业研究单位如课题组等使用。CD—ROM光盘具有数据稳定可靠，使用方便的特点。在目前我国数据库使用主要靠微机进行的情况下，销售光盘数据库已成为一般数据库服务的主要方式，其主要用户为各大专院校、综合性图书情报单位及企业集团。我们推出的CD—ROM光盘受到用户的欢迎，已有一些科研院所、大专院校联系购买。

网络版数据库只有通过国内外联网，实现资源共享，才能不断发展壮大。而实现数据库可持续发展，联网具有双向选择的特点，它利于用户选择数据库，也有利于数据库的资源更新，提高数据库为社会共享的能力和利用率。目前，我们已申请设立新课题研究网络版的数据库。

总之，只有加强数据库全面的质量控制，才能建立一个高品质的专业数据库。在研制期间与试运行时，我们分别邀请了一些专家教授及科技人员现场操作使用，都觉得收录范围较广，使用方便，查询快捷，科学实用，是科学工作者进行科学研究的好帮手。

参 考 文 献

1　刘强．关于加强数据库建设的思考．情报科学技术，1996；(1)

2　谭显华，宁继珍．中文科技期刊数据库商品化尝试．情报学报，1997；(5)

（责编：柽钧京）