

●纪蔚蔚 潘有能

文献数据库质量控制系统的实现^{*}

摘 要 文献数据库质量与建库成本、建库周期之间存在着矛盾。通过数据录入中的质量控制、主题词标引及分类中的质量控制、数据整合中的质量控制,可以解决这一问题。图 1。参考文献 2。

关键词 数据库 质量控制 文献处理

分类号 G250.74

ABSTRACT There is a contradiction between the quality of a document database and the creation cycle and cost for the database. In this paper, the authors think that we can solve some problems by quality control in the inputting, subject indexing, classification and data integration. 5 figs. 2 refs.

KEY WORDS Database. Quality control. Document processing.

CLASS NUMBER G250.74

1 质量控制的提出

文献数据库是提供用户服务、制作光盘检索系统、进行因特网上查询和文献计量的基础,也是进行文献自动化处理的素材。数据库质量有问题,便会造成文献在数据库中的漏检或误检,还会造成自动标引的准确率低,甚至无法实

现,文献计量指标及统计分析的可信度也不能得到保障。质量,是数据库的生命,它直接影响用户服务和文献利用,这已经成为人们的共识。

文献数据库的质量主要体现在:数据录入的准确性,数据收集的完整性,数据更新的及时性,数据处理的正确性。常见的错误类型有:字段类型不规范,相关字段不匹配,关

标准,是一个运行在 TCP/IP 协议之上的应用层协议,它规定了客户机查询服务器以及提取结果记录等过程中所涉及的数据结构和数据交换规则,从而解决了现存书目数据库检索接口的异构性问题。相对于 OAI, Z39.50 的功能更加完善,但也带来实现的困难和费用的高昂。一般说,只有标引详细,数据质量很高,对互操作质量要求相当苛刻的系统才采用。

4.3 资源描述框架 RDF 与可扩展标记语言 XML

由 W3C 推出的 RDF 是一套描述资源及其属性和属性值的模型,其制定的目的主要是为元数据在 Web 上的应用提供一个基础结构,以方便不同元数据间的互操作。可扩展标记语言 XML 作为元数据的编码标准,提供了元数据在语法层次上的互通性,使它跨越特定平台、特定系统的限制。使用 RDF/XML 命名域的概念,在创建一个元数据格式时,借用其他元数据集的某些元素,可以减少重复劳动并增强元数据格式间语义互通性,方便互操作的实现。

元数据是数字图书馆建设的关键技术之一。数字化图书馆的运作,无论是数据的加工、存取,信息的浏览、检索,还是资源的整合与长期保存都是以元

数据为基础实现的。随着数字图书馆的发展,元数据的研究必将进一步深入。

参考文献

- 1 The Dublin Core Metadata Element Set. ANSI/NISO Z39.85-2001
- 2 Marcia Lei Zeng. Supporting metadata interoperability: trends and issues. Proceedings of 21st NIT international conference. Beijing: Tsinghua University Press, 2001
- 3 林海青. 数字化图书馆的元数据体系. 中国图书馆学报, 2000, 26(4)
- 4 吴政. 都柏林核心集在图书馆应扮演的角色. 上海中文元数据应用国际研讨会, 2001

吴开华 清华大学图书馆研究馆员。通讯地址:北京清华大学。邮编 100084。

邢春晓 清华大学计算机系副教授, 博士后。通讯地址同上。

罗德胤 清华大学建筑学院博士研究生。通讯地址同上。

(来稿时间:2001-10-29)

^{*} 本文系江苏公安专科学校科研项目《公安文献全文数据库及计算机辅助标引与检索系统》(97XB870001)的成果之一。

联词库不一致,字段内容不统一及错漏等。

文献数据库的质量问题已得到人们的普遍关注,但数据库质量与建库成本、建库周期之间却存在矛盾。在自动化水平较低的数据库系统中,要在减少并校改上述错误的同时保证建库速度,必然要以巨大的人力、物力为代价,这是数据库建设者普遍难以承受的。为此,我们研制了文献数据库质量控制系统,并投入全国公安院校图书馆联合共建公安文献全文数据库使用。实践证明,该系统能显著地提高建库效率,确保建库质量,缩短建库周期,降低建库成本,初步实现了数据库质量控制的自动化和智能化。

2 质量控制的实现

质量控制问题贯穿整个数据库建设的始终,在数据库系统设计阶段,就应考虑质量控制问题。一般来说,对数据质量和建库速度产生影响的环节主要有以下几个:数据录入、主题词标引及分类、数据整合。我们尽量采取计算机自动化、智能化技术,辅之人工机助的方式,建立一个比较完善的质量控制体系。

2.1 数据录改中的质量控制

为了对文献进行更有效的管理和质量控制,我们将非结构化的源文献转换成结构化数据。在数据库中,可以方便快捷地根据各种需要对字段进行索引、排序、整理、统计等工作,以利于对错误信息的甄别和校改,提高文献的有序度及准确率。

在实际应用中,将非结构化源文献转化到结构化数据库中的途径主要有两种:人工录入及自动录入。人工录入不但效率低,费用高,而且会造成漏输、误输、重输等各种各样的错误,而自动录入既减轻了著录人员输入汉字的负担,又减少了人工判断错误,既防止著录项目遗漏,又确保了著录项之间的正确关联,能大大加快数据录入速度,提高录入数据的正确率。

在进行自动录入前,我们必须对文章进行标记。为了提高标记的速度及准确率,可以利用代码字典进行一些前处理。录入过程中,再利用代码字典的对应关系,将字段的实际内容转换到文献数据库中,确保数据录入的准确性、规范性和一致性。

2.2 主题词标引及分类中的质量控制

人工语言与自然语言相结合的检索系统能确保文献数据库的检索质量。标引质量是人工语言处理中最重要的一个影响因素,而文献主题词标引与文献分类一直被视为最主要、最常用的标引途径。传统的人工标引工作效率不高、一致性低,而自动标引标引速度快、一致性好、规范性强。但自动标引也有不足,如字面硬匹配、隐含主题概念不易表达等。为了提高自动标引的质量,我们对传统的词典分词法和词表分类法作一些改进,将词表建设和自动分词、分类融为一体,这样不但可以达到较好的标引效果,还可以不断

完善系统词典,解决了词典分词法中词典的构造困难、更新滞后问题,提高了主题词标引的质量和效率,并能通过对词表分类法的完善与优化,使自动分类的结果达到实用水平。

为确保文献标引质量和词表的自学习,必须对标引生成的新词和模糊词进行控制。具体可以采取人工机助的方式进行,系统将标引生成的新词和模糊词以及相关的词频、相似词、文献记录号等信息提交给标引员进行人工维护,根据维护结果,系统自动更正数据库中对应的错误信息,同时被相关词表所学习。这样,标引数据能达到相当高的一致性和准确性,并能使词表不断地进行自学习,最后趋于稳定,整个系统得到完善。

2.3 数据整合中的质量控制

数据整合中的质量控制是将各录改站点的信息上载到主数据库中,对合并后的主数据库进行整序、衍生,利用衍生数据库(关联字典库)对主数据库进行规范控制的数据处理。

在文献数据库整合中,经常需要对库中记录的相关信息做一些修改,例如标引词之间的一致性处理,分类号的修正,作者名及机构名称的更改等。假如单纯的由人工在数据库中直接进行修改,不仅增加了工作量,而且容易造成漏改、多处更改不一致等错误,以至影响文献数据库的质量。为了更好地实现质量控制,给数据库维护人员提供一个方便、快捷、友好的对数据库记录进行修改的途径,我们可以建立一些关联字典,如分类关联字典、主题词关联字典、机构关联字典、作者关联字典、期刊关联字典等。关联字典和文献数据库中的记录建立有机联系,以达到对数据库进行自动更改的目的。

关联字典由文献数据库动态生成,以人工机助的方式进行维护。关联字典设有规范词、非规范词、关联项、文献记录号、词频等字段。其中“关联项”主要记录规范词与非规范词之间的关系,如“用”、“删除”、“用后删除”等;“文献记录号”则提供了关联字典和文献数据库中记录的联系途径。数据库维护人员可以根据系统提供的词频、是否新词、相似词等信息对关联字典进行维护,同时,这种修改将自动反映在文献数据库中。关联字典的应用可以将相同的错误全部一致修正,大大减轻数据库维护人员的工作量,减少维护过程中可能产生的错误,使文献数据库的质量得到保证。

我们还可以建立一些代码字典,用来存放系统中使用到的代码型数据,不但可以减少数据库的冗余度,提高数据库的存储效率,还可以使某些特定字段的值得到规范控制,减少数据库维护的难度。一般的代码字典有期刊代码字典、地区代码字典、学科代码字典等。代码字典的主要字段有:代码、代码类型、代码名称等。

3 质量控制的数据流

文献数据库质量控制系统的每一个功能均是紧紧围绕

文献数据而展开的,科学的文献处理作业流程是数据库高质量与高效率的保证。图 1 给出了质量控制系统的数据库,从图中可以看出从源文献导入开始,经过自动分词、自动分类、数据整合、词表学习、关联字典生成与维护,直到文献数据库规范控制的整个数据质量控制过程。

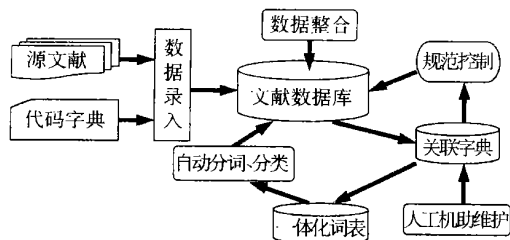


图 1 文献数据库质量控制数据流程图

4 《公安文献全文数据库》的质量控制

文献数据库的质量控制是一项系统工程,一个高质量的数据库主要表现在其数据齐全、数据准确、数据更新及时、检索系统方便快捷等方面。它的建成不仅要有一定的技术保障,还需要有力的组织保障。全国公安院校图书馆在联合共建《公安文献全文数据库》时不但采用“文献数据库质量控制系统”,还采取了一系列相应的组织管理措施,使质量控制收到了良好效果。

4.1 组织保障措施

公安文献全文数据库质量控制系统,是江苏公安专科学校科研项目《公安文献全文数据库及计算机辅助标引与检索系统》的一个子系统。该项研究受到了江苏省高校图书馆工作委员会的高度重视,并将该项目作为江苏省高校文献信息保障系统的一个子系统予以立项资助。它的研究成功,促成全国 17 所公安院校图书馆联合共建公安文献全文数据库。联合共建工作由全国公安院校图书馆协作组主席馆主持,江苏公安专科学校图书馆提供技术支持并承担数据汇总、质量控制、制作检索系统光盘和网上查询等工作。共建馆每年召开一至两次共建工作研讨会,制定数据库共建章程与协议,集中讨论、解决一些实际工作中可能遇到的问题。公安文献全文数据库现已收录我国 300 多种公安专业期刊 1997 年至 2001 年中 6 万多篇文献信息。同时建成了具有一定准确性与完备性的各类规范字典、公安分类表、公安主题词表等。

4.2 确定建库目标

我们对公安文献全文数据库的专业范围、数据库类型、数据来源等进行了广泛调研和仔细论证,以确定建库目标,树立数据库特色,确保数据库质量。

公安文献全文数据库的建设目标为:构建公安文献知识库体系,全面、客观、系统地反映我国公安专业期刊文献;实现文献处理自动化、智能化;提供多途径、多角度、全方位的查询手段,使用户高效、方便、快捷地获取公安文献信息;提供我国公安科学研究成果统计、分析与评价依据。

4.3 选定与划分文献源

为实现建库目标,文献源的选定是极为重要的环节,它

不但直接影响文献的查准率和查全率,而且关系到数据库的权威性和代表性。文献源选定与划分时,主要考虑学科在国内外专业分布中的位置,学科体系以及学科的独立性和相互交叉等因素。公安文献全文数据库的主要学科有:警察学、治安管理学、侦查学、侦查技术学、消防管理学、交通管理学、犯罪学、罪犯学等。数据来源以我国公安期刊为主,数据库类型为全文文献型。由此,公安文献全文数据库的规模与范围便基本确定,在此基础上,我们制定了文献源的选定原则及步骤,并对共建任务进行分配。任务划分时要充分考虑到文献源的地区分布情况、共建馆的典藏特色、人力、物力资源情况等。

4.4 制定标引与数据制备细则

标引细则直接关系到数据质量、开发深度以至整个系统的功能。其主要内容包括:文献的标引项目、标引格式、标引模式、标引方法、标引使用工具等。公安文献全文数据库的标引项目有:篇名、作者、机构、刊名、卷期、页码、关键词、主题词、分类号、摘要、正文、引文等。文献分类标引采用《公安文献分类法》,主题标引采用系统自建的“分类号—主题词—关键词—一体化词表”,以“自动标引后控制”的方法进行。

共建数据在制备过程中有明确的质量要求,电子文本需经过两次以上的校改,一是印刷型文档通过电子扫描、OCR 汉字识别转换后,利用其编辑软件校改;二是利用专业智能校改软件进行校改,同时要对本文进行必要的标记处理。

4.5 规定数据传输期限与方式

为保证数据库的时效性,必须及时增加新数据。共建工作规定各共建馆必须定时、定量将数据传送至汇总馆。数据通过电子邮件的方式进行传输,以压缩形式加入电子邮件的附件。电子邮件的主题用规范格式标明发件单位、数据内容及发送批数等信息,以确保数据传输的安全、及时与有效管理。

4.6 数据汇总与质量控制

收到数据后,数据汇总馆即利用“文献数据库质量控制系统”进行由“数据导入”开始的一系列文献处理与质量控制工作。利用自行开发的“共建数据管理系统”,将质量控制工作中发现和处理的 all 问题以日志形式一一详细记载,并及时传达各共建馆。管理系统还对各共建馆完成处理数据的数量、时间、质量评分、缺刊情况等进统计与评价,为共建工作提供决策参考和管理依据。

参考文献

- 1 李勇,安博,张昱.加强文献数据库及检索刊物的质量控制.情报学报,1999(1)
- 2 曾海燕.建立文献数据库规范控制的探讨.现代图书情报技术,1997(1)

纪蔚蔚 女,江苏公安专科学校图书馆馆员。通讯地址:江苏南京。邮编 210012。

潘有能 南京大学信息管理系博士研究生。通讯地址:江苏南京。邮编 210093。

(来稿时间:2001-11-16)