第16卷 第6期
1997年12月

情 报 学 报

Vol.16, № 6
December, 1997

书目数据质量控制研究

李 敏
(中国科学院文献情报中心,北京 100080)

摘要 文章分析影响书目数据质量的因素,利用层次分析法计算这些因素的重要性排序,据此提出加强书目数据质量控制的措施。

关键词 书目数据 质量控制 层次分析法

Study on the Quality Control in Bibliographic Data

Li Min

(*Documentation and Information Centre of the Chinese Academy of Sciences, Beijing* 100080)

Abstract This paper analyses the factors of influencing data quality and refers to an Analytic Hierarchy Process to find the way of increasing bibliographic data quality.

Keywords bibliographic data, quality control, Analytic Hierarchy Process.

书目数据库是图书馆自动化建设的核心,其质量的好坏直接关系到整个信息管理系统的生存和发展,关系到文献数据库产品的命运。因此,对书目数据的质量进行控制与评价有着十分重要的意义。本文结合我国书目数据建设的实际情况,对书目数据质量控制与评价方法提出一些想法。

1 中外书目数据质量控制状况

改革开放以来,我国数据库建设成绩斐然。到1995年底,我国自建数据库总数已达1038个,其中书目数据库将近100个,累积文献记录量100多万条,其数量逐年增加,已达到相当规模。但同时也存在一些问题,最突出的是普遍重视数量的增长,而对数据质量要求不严。主要问题表现在:重复数据太多;著录内容太少;标目不一致和录入错误太多;机读数据未严格遵循有关标准和规则;对数据错误缺乏分析研究;对数据生产缺乏定性和定量分析;缺乏较高层次的组织管理等。相比之下,国外的书目情报机构不但重视数据数量,而且将提高数据质量作为系统发展的长期目标,投入了大量的人力、物力进行研究,产生了许多成功的质量控制方法。如(1)OCLC建立了数据质量控制部,加强数据质量软件的开发、编目人员的培训和反馈控制;(2)SOLINET LAMBDA系统成立了数据质量委员会,强调遵循编目条例和复用LC的记录;(3)Richard Reeb提出在数据质量控制中采用统计质量管理方法;(4)Dorothy Mcpherson通过

收稿日期:1997年10月5日

作者简介:李 敏,女,1966年生,讲师,中国科学院文献情报中心研究生。

分析加州大学 MELVYL 系统中影响数据质量的因素，提出了相应的质量控制策略；(5) E.J.Yaunakovadakis 等研究了数据质量控制的专家系统；(6) C.E.Welse 等提出了应用 ISO9000 系列标准保障数据的质量。这些质量控制措施在保证机读数据的质量方面发挥了重要的作用。

2 书目数据质量的影响因素

影响书目数据质量的因素有：

(1)文献原文的质量是否符合要求。

(2)各种标准、规范是否齐全和严格，数据是否完整、统一、准确。这包括机读数据的格式标准；著录规则；分类表、主题词表以及相应的标引规则；著录项目、检索点的完整性等。

(3)工作人员的水平和素质。这包括人员的专业知识和文献业务水平，录入人员的水平与素质。一般说来，一个称职的机读编目人员应具备：(1)一门或几门专业知识，(2)计算机知识，(3)图书情报知识，(4)还要具备现代思想观念和敬业精神以及良好的心理素质。

(4)工作流程是否合理。一个完整的书目数据生产流程应包括：数据查重→制作工作单→核对→录入计算机→校对→数据总复查→数据成形并交送用户→反馈控制→参照系统的维护。

(5)组织管理的科学性。现代管理强调满足从业人员的情感和成就感，注重通过激励和提供参与管理的机会来提高生产率，让每一个成员参与管理过程，组织重大决策。组织管理者既要协调好编目人员、编目软件和计算机设备，又要制定出科学的管理方法和制度。

(6)软件系统的准确与完善。编目软件系统只有性能好、功能全，才能够生产出高质量的机读目录数据。

(7)计算机质量控制功能的覆盖面和深度。在编目工作中，计算机本身的性能直接影响着图书查重、数据录入、数据转录等工作的效率和速度，有时由于计算机的原因会在数据制作过程中出现非人为的错误，为此，要考虑到设备更新问题。

3 各影响因素重要性评价

以上因素对书目数据质量的影响程度是不同的，下面用层次分析法对各影响因素的相对重要性进行评价，并在此基础上提出加强书目数据质量控制的措施。

层次分析法的核心是把复杂的决策问题分解成若干层次，根据问题的性质和要达到的目标，再把同一层次的问题分解为不同要素，并按要素间的相互关联及隶属关系，形成一个多层次、多因素的分析结构模型。对本文所要解决的问题来说，最终归结为最低层次各因素相对于最高层次的相对重要性的排序问题。具体步骤如下：

3.1 层次分析模型的构造

我们的总目标是：提高书目数据工作质量。影响总目标实现的准则层包括：文献原文质量，文献前处理质量，数据录入建库的质量。通过对直接影响准则层的各因素进行分析，最后选定六个因素为措施层。层次模型如图 1 所示。

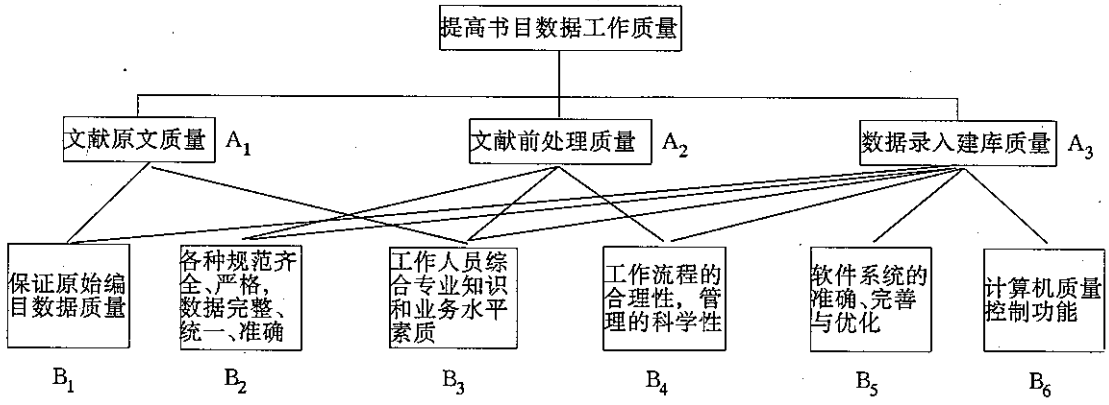

图 1　书目数据质量控制的层次模型

3.2　判断矩阵及其标度

在层次分析法中，用判断矩阵计算某一层的各因素相对于上一层某因素的重要性权重。判断矩阵中两两因素的比较，采用 T.L.Saaty 提出的标度方法，其含义列于表 1。

表 1　判断矩阵标度及其含义

标度	含　义
1	表示两个因素相比，具有同样重要性
3	表示两个因素相比，一个因素比另一个因素稍微重要
5	表示两个因素相比，一个因素比另一个因素明显重要
7	表示两个因素相比，一个因素比另一个因素强烈重要
9	表示两个因素相比，一个因素比另一个因素极端重要
2,4 6,8	上述两相邻判断的值
倒数	因素 i 与 j 比较得判断 b_{ij}，则因素 j 比 i 比较的判断 $b_{ji}=1/b_{ij}$

经过分析讨论构造以下四个矩阵：

O	A_1	A_2	A_3
A_1	1	2	4
A_2	$\frac{1}{2}$	1	3
A_3	$\frac{1}{4}$	$\frac{1}{3}$	1

O-A 矩阵

A_1	B_1	B_3
B_1	1	1
B_3	1	1

A_1-B 矩阵

A_2	B_2	B_3	B_4
B_2	1	$\frac{1}{5}$	$\frac{1}{3}$
B_3	5	1	3
B_4	3	$\frac{1}{3}$	1

A_2-B 矩阵

A_3	B_1	B_2	B_3	B_4	B_5	B_6
B_1	1	$\frac{1}{5}$	$\frac{1}{4}$	1/2	1/2	$\frac{1}{3}$
B_2	5	1	2	4	3	5
B_3	4	$\frac{1}{2}$	1	3	3	3
B_4	2	1/4	1/3	1	2	2
B_5	2	1/3	1/3	1/2	1	1/3
B_6	3	1/5	1/3	1/2	3	1

A_3-B 矩阵

3.3 层次单排序及一致性检验

层次单排序是指在层次分析中，由单一判断矩阵各因素之间对其准则的重要性权重排序，可通过计算机判断矩阵的最大特征根所对应的向量来确定。计算方法采用和积法。以 O-A 矩阵为例介绍其计算过程。

(1)将判断矩阵每一列正规化：

$$\bar{a}_{ij}=\frac{a_{ij}}{\sum_{k=1}^{n}a_{kj}} \qquad i,j=1,2,\cdots,n$$

经过列正规化以后，O-A 判断矩阵为：

O	A_1	A_2	A_3
A_1	0.5714	0.6000	0.5000
A_2	0.2857	0.3000	0.3750
A_3	0.1429	0.1000	0.1250

(2)对每一列经正规化后的判断短阵按行相加：

$$\bar{w}=\sum_{j=1}^{n}a_{kj} \quad i,j=1,2,\cdots,n \quad 得：\quad \bar{w}=\begin{bmatrix}1,6714\\0.9607\\0.3679\end{bmatrix}$$

(3)对向量 $\bar{w}=[\bar{w}_1,\bar{w}_2,\cdots,\bar{w}_n]^T$ 正规化

$$wi=\frac{\bar{w}_i}{\sum_{j=1}^{n}\bar{w}_j} \qquad i,j=1,2,\cdots,n$$

得所求特征向量：$w=[0.5511,0.3203,0.1226]^T$

(4)计算判断矩阵最大特征根

$$\lambda max=\sum_{i=1}^{n}\frac{(AW)_i}{nw_i}$$

式中 A——判断矩阵，w_i——判断矩阵特征向量分量

$$AW=\begin{bmatrix}1 & 2 & 4\\1/2 & 1 & 3\\\frac{1}{4} & \frac{1}{3} & 1\end{bmatrix}\begin{bmatrix}0.5571\\0.3202\\0.1226\end{bmatrix}=\begin{bmatrix}1.6881\\0.9667\\0.3686\end{bmatrix}$$

$$\lambda max=\frac{1.6881}{3\times0.5571}+\frac{0.9667}{3\times0.3202}+\frac{0.3686}{3\times0.1226}=3.0183$$

(5)判断矩阵的一致性检验

根据最大特征根,先计算偏离一致性指标 CI。由公式 $CI=\frac{\lambda max-n}{h-1}$,然后将 CI 值与相应阶数矩阵的平均随机一致性指标 RI(见表 4)相比,$CR=\frac{CI}{RI}$。CR 为称随机一致性比率。当 CR <0.10 时,即认为判断矩阵具有满意的一致性,可做层次分析。否则,需要调整判断矩阵,直到具有满意的一致性为止。

n	1	2	3	4	5	6	7	8	9
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

本例计算如下:

$$CI=\frac{3.0183-3}{3-1}=0.0092$$

$$CR=\frac{CI}{RI}=\frac{0.0092}{0.58}=0.0159<0.10$$

证明 O－A 判断矩阵具有满意的一致性。同理可计算出其他三个判断矩阵的特征向量 w_1,λmax、CI 和 RI 值。

A_1－B 矩阵:w=(0.5000, 0.5000), λmax=2, $CI_1=0$, $RI_1=0.0000$, $CR_1=0.0000$。

A_2－B 矩阵:w=(0.3200, 1.9410, 0.7850), λmax=3.0360, $CI_2=0.0180$, $RI_2=0.58$, $CR_2=0.0310$。

A_3－B 矩阵:w=(0.0532, 0.372, 0.2475, 0.1222, 0.0819, 0.1190), λmax=6.3750, $CI_3=0.0750$, $RI_3=1.24$, $CR_3=0.0604$。

以上计算结果说明 A_i－B 矩阵均满足一致性。

3.4　层次总排序及其一致性检验

层次总排序是综合层次单排序的结果。用上一层次因素的组合权值加权,即得到下一层因素相对于上层整个层次的组合权值。逐层计算,直至排出最低层各因素对最高层(总目标)的相对重要性权值,即相对重要性的排序值(表 3)。

表 3　层次总排序

上层 A / 下层 B	A_1	A_2	A_3	层次总排序	序　号
	0.5571	0.3202	0.1226		
B_1	0.5000	0	0.0532	0.2850	2
B_2	0	0.3200	0.3762	0.1486	4
B_3	0.5000	1.9410	0.2475	0.9000	1
B_4	0	0.7850	0.1222	0.2663	3
B_5	0	0	0.0819	0.0100	6
B_6	0	0	0.1190	0.0146	5

$$CI_{总}=\sum_{i=1}^{n}a_i(CI_i)=0.5571\times0+0.3203\times0.018+0.1226\times0.0750=0.01496$$

$$RI_{总}=\sum_{i=1}^{n}a_i(RI_i)=0.5571\times0+0.3203\times0.58+0.1226\times1.24=0.3374$$

$$CR_{总}=\frac{CI_{总}}{RI_{总}}=\frac{0.01496}{0.3374}=0.04433<0.10$$

以上得出的层次总排序满足一致性要求,其排序结果是:

$$B_3\rightarrow B_1\rightarrow B_4\rightarrow B_2\rightarrow B_6\rightarrow B_5$$

以上排序结果,为制定书目数据质量控制措施提供了理论和实际相结合的可靠依据。

4 加强书目数据质量控制的措施

质量控制是一项系统而复杂的工程,综上所述评价方法,应主要采取以下措施。

4.1 提高人员素质,制定全员培训计划。

从事书目数据工作的人员应具备以下素质,应按这些素质要求进行全员培训。

(1)掌握现代编目理论,有较深的图书情报学专业知识。可对编目人员特别是非图书情报专业毕业的编目人员不定期地举办编目理论学习班,聘请资历较深的编目专家和学者授课。

(2)尽可能熟悉各学科知识。编目人员面对的是加工各学科、各专业的图书资料,如果知识面窄就难以胜任工作。

(3)具备计算机操作能力。现代编目工作均应用计算机技术,不懂得计算机知识,缺乏运用和操作计算机的能力,就会大大影响书目数据的编制工作。

(4)具备相当程度的外语水平。编制中文图书的机读目录,外语知识日显重要;编制西文图书机读目录,外语知识则是必备工具。

(5)具有职业意识和敬业精神。一个具有职业意识和高度敬业精神的组织群体是实施和保证数据质量的关键,应把此作为管理层的一项奋斗目标。

4.2 正确使用原始文献,提高文献源质量

4.3 加强编目工作科学化管理和数据生产的定量化管理

良好的编目工作管理制度是保证机读书目数据质量的关键。这包括对编目人员的培养、教育和合理使用;对计算机等技术资源的科学利用与开发;以及对整个编目工作流程的科学控制。这是一个产生质量和效益的过程。编目工作的科学化管理,要从定性和定量两个方面着手,尤其是要重视数据生产的定量化管理。编目数据可分成三级:一级为完全级,表示该记录建立时依据了原作品,著录项目齐全;二级为次完全级,表示该记录建立时未依据原作品,仅根据编目卡片,著录项目较齐全;三级为不完全级,表示该记录建立时仅根据编目卡片,著录项目不齐全,缺一些重要字段。级别不同,生产费用不同,产生效果也不同。若进行适中的定量分级,将会调动数据生产人员的积极性,保证数据质量和数量的同步提高。定量管理贯穿于整个书目数据加工过程中,要形成一套科学的方法并成为制度。

4.4 加强对各种标准、规范的理解和执行

目前,我国许多图书馆所生产的编目数据不都是可用来交换的机读书目数据。只有严格按照各种规范、标准进行著录、标引,才能产生完整的、统一的、准确的高质量机读书目数据。

为此，我们应加深对各种标准，规范的理解，尽量避免由于对格式、标准、规范掌握不准而导致的人为错误。

4.5 不断调整结构，提高和优化系统功能

随着书目数据编制工作的发展和深化，会出现许多新问题和新要求，原有的系统功能无法满足需要。这就要求在自动化的环境下，围绕数据生产、管理、维护和使用全面进行结构调整，对编目系统进行优化和更新。软件系统功能越完善、性能越优，流程越合理，就越能保证书目数据的详细、全面、有效和完整。与此同时，要经常对数据库进行维护优化，更新计算机设备，利用技术上更先进的计算机系统。

参 考 文 献

[1] 黄箭，张建勇：机读书目数据质量控制研究——提高 TOTALS 系统中书目数据质量的策略，《现代图书情报技术》，1997，(增刊)
[2] 李正祥：论机读书目数据的质量控制，《图书情报工作》，1997，(5)
[3] 万述鸿：论书目数据库标引质量的评价与控制，《现代图书情报技术》，1991，(2)
[4] 黄国俊，张京：文献数据库质量的计算机控制，《现代图书情报技术》，1990，(1)
[5] 张建勇：关于机读书目数据生产的质量保障体系的研究，中国科学院文献情报中心硕士研究生毕业论文，1994
[6] 刘全根：《科技情报分析研究》，甘肃，甘肃科技出版社，1993

(责任编辑 许增棋)

中国科技情报学会竞争情报分会'97年会召开

中国科技情报学会竞争情报分会以"企业信息化与竞争情报咨询服务"为主题的'97年会9月21日至26日在浙江富阳召开。来自全国的82名竞争情报和信息咨询工作者出席了会议，收到论文77篇。与会代表围绕会议主题开展了热烈的交流和讨论，就以下问题取得了共识：(1)随着我国社会主义市场经济的发展，尤其是15大提出了实行国有企业股份制改造的方针，企业竞争将日益加剧，这就为以增强国家、地区和企业竞争力为目的、以市场竞争为内容、以竞争对手为核心的我国竞争情报工作创造了良好的社会契机和政策环境，鼓舞和增强了代表们将竞争情报理论与中国市场经济实践结合起来、积极开拓和搞好我国竞争情报工作的信心和决心。

(2)几年来，我国竞争情报工作取得了很大的发展，竞争情报咨询工作已经成为我国信息和咨询部门的重要工作之一。随着一些竞争情报学术论著的出版，大学关于竞争情报专业课程的开设，一批竞争分析方法的应用，一些国家机构和地方政府对竞争情报投入的增大，我国竞争情报工作呈现了良好的发展势头。

(3)竞争情报在我国还处于初创阶段，应当实行官产学相结合，紧紧抓住提高国家、地区、集团和企业竞争力这个中心议题、加大竞争情报工作的宣传力度；继续搞好分会的组织建设，拓宽分会的组织基础，发挥分会的群体作用；加强竞争情报的理论与方法建设，加强普及培训工作，逐步形成网络式的竞争情报培训中心；加大竞争情报实践活动，作出具体的成绩和贡献，从而提高竞争情报和竞争情报分会在全国的影响力。

(4)办好明年分会成立三周年的纪念活动，出版专辑，组团参加"竞争情报专家协会"第13届国际年会，邀请外国专家来华讲学等。