

数据制作视角下的书目数据分析

原小玲

(太原科技大学图书馆, 太原 030024)

摘要: 本文从数据制作流程角度入手, 发现中文图书书目数据在完整性、准确性与可共享性上都存在问题。通过对数据制作流程的分析, 从数据制作视角对这些问题的成因进行归纳, 进而提出建立以图书在版编目数据为源头的全国统一书目数据, 并在出版社及图书下游环节间建立数据通道, 重视名称规范库建设与推广使用等对策, 以期对因数据制作流程产生的数据问题的改进有所裨益。

关键词: 中文图书; 书目数据; 数据质量

中图分类号: G254

DOI: 10.3772/j.issn.1673-2286.2018.11.005

书目数据在揭示图书信息、进行文献组织、帮助用户检索图书、实现资源共享与图书聚类分析等方面有不可替代的作用。截至2018年8月, 国家图书馆的全国图书馆联合编目中心拥有中文书目数据超过1 300万条^[1]; 截至2018年3月, 中国高等教育文献保障系统 (China Academic Library & Information System, CALIS) 联合目录中心网站显示, 其数据库中的中文数据已有413万多条^[2]。书目数据的质量很大程度上取决于数据制作过程中各参与主体的表现。目前的书目数据制作由图书在版编目数据 (Cataloguing in Publication, CIP) 中心、公共图书馆系统、高校图书馆系统3大主体完成, 依据不同标准生成数据的结构, 详略等均有不同。分析其数据制作流程, 发现其中的问题与不足, 整合资源, 在数据制作环节建立全国统一的书目数据, 各使用方再根据实际对其进行减法操作与少量修改, 并增加馆藏等个性化信息, 以适应自身管理需求与馆际数据交换需要。这样既有利于数据交换与共享, 又可以极大地节约社会成本, 还能真实反应出版量与图书馆入藏量的关系。

1 数据制作流程

数据制作就是以图书为加工对象, 根据一定的规则, 对图书的形式与内容特征进行揭示, 生成书目数据

的过程。从数据制作主体及其产品来看, 中文图书书目数据目前由2个环节的3个主要主体制作, 并使用各自标准, 制作流程也有所区别。

CIP数据制作流程是在出版社发稿付排前, 由责任编辑按《图书在版编目工作手册》要求, 填写“图书在版编目工作单”, 将工作单提交版本图书馆, 版本图书馆采用《图书在版编目》数据标准, 编制CIP数据, 标引中图分类号与主题词, 并分配给该CIP数据一个审定号, 加工成CIP数据标准格式后返回出版社, 出版社将其印制于图书版权页。《图书在版编目》数据标准与《普通图书著录规则》国家标准保持一致, 但不采用任何国际标准。

图书进入流通环节后, 各图书馆为组织文献与服务用户制作图书编目数据。编目数据是由各馆分类编目人员共同完成。他们依据到馆图书的外在特征与内容特征, 对图书进行全面揭示, 并依编目标准对著录与标引项进行规范。按照编目规则生成MARC数据后, 发布于本馆OPAC系统, 供资产管理与用户检索所用。国家图书馆与CALIS分别建立有联编中心, 各成员馆可以将数据上传, 经中心审核后, 如为重复数据, 只增加馆藏项; 如无重复数据, 则列为新增数据, 后续提交该数据的馆只增加馆藏项。为保证数据质量, 各联编中心都给有权限的成员修改数据的权限, 以不断完善数据。但以国家图书馆为代表的公共图书馆采用《中国文献

编目规则》(以下称“国图标准”),以CALIS为代表的高校与研究型图书馆采用《CALIS中文图书编目规则》(以下称“CALIS标准”),这两个标准都遵循国际书目标准ISBD制定,但有细微差别。

由于同一种书由3个部门依据不同标准生成书目数据,造成数据数量远超过图书种数的现状。而且数据的制作标准不同,生成结构与详略程度都不同的数据,虽然满足不同需求,但其冗余使基于数据的服务缺乏精准性。

2 书目数据质量现存问题分析

2.1 数据完整性低

书目数据完整性指数据对图书信息揭示的全面性,具体指规定的可见信息源上的项目著录完整,标引的信息内容揭示角度全面,以及其他相关信息尽可能揭示完备。目前使用的3种书目数据在完整性上有不同程度的欠缺。

(1) CIP数据完整性要求低。图书生产环节的CIP数据,由图书形式与内容生产者参与,最有条件做到数据完整。但其最大问题恰是描述项目过于简单。传统的CIP书目详细信息中有正题名、ISBN、版次、印次、价格、开本、装帧、分类号、主题词、内容提要等项,但缺失责任者项,且正文语种项为空,书内图表信息也不予揭示。尤其是责任者项与语种项缺失,对于下游环节来说,该书目数据的实用价值降低很多。《中国出版物在线信息交换图书产品信息格式规范》(CNOIX)标准为有效传递图书产品信息到产业链各环节,对包括图书内容、出版、营销等有关信息在内的图书产品交易信息的XML格式做了具体要求,项目较传统CIP更详细。但对于产品价格、语种、货币、制造国及套书信息等书目数据中的重要信息,该标准规定为可选而非必选^[3],导致数据在完整性上与传统CIP数据要求相比,仍然没有突破性提升。

(2) 不同层级编目数据完整性不同。编目数据的完整性问题体现在它们对书目记录级次的区分上。根据书目信息或内容标识的完整程度,国图数据将书目记录分为4个等级,即完全级与次级1~次级3。其中,完全级为最完整的MARC记录,编制记录时与编目实体核对过;次级1为完整记录,但未核实文献实体;次级2指CIP记录,它通常是不完整的;次级3为不完整编目的记

录,且书目发行机构一般不对其进行升级。CALIS数据在国图分级基础上增加定义了次级4,通常指采访级简编记录^[4]。从完全级到次级1~次级4的书目记录,其详细程度逐级降低,也就表明数据的完整性逐级降低。编目数据的完整性问题在对必备字段的规定上也有体现。国图标准与CALIS标准对字段必备性要求不同,总体上CALIS标准较国图标准更为全面、严格。国图标准中推荐与选择使用的字段,在CALIS标准中是有且必备,而非完全的自由选择使用。对字段使用的不同要求,造成数据详略程度不同。数据级次越低,其完整性就越差。

2.2 数据准确性差

不论是CIP数据还是编目数据,都是对实体图书的描述与揭示。数据准确一是体现在对图书外在信息的准确著录,“有书在手”才能真正做到著录准确;二是体现在对图书内容的准确分析与标引,对图书内容了解才能保证分析与标引的准确。

由CIP数据制作流程可见,出版社只是向出版总署提交一份工作单,既未附带其他校样,也未由图书作者帮助标引相关主题词^[5],因而CIP数据分类与主题标引都很难准确,“目前CIP数据的准确性离社会各界的要求尚有一定的差距”^[6],是图书馆界对其诟病最多的一个方面。

国图数据与CALIS数据实行成员馆共建共享,可以做到数据制作时“有书在手”,相对CIP数据而言,准确性要高。尤其是完全级数据,要求与编目实体进行核对。但其他次级数据不要求与编目实体进行核对,当不与编目实体进行核对时,数据准确性是没有保障的。在CALIS与国图的数据中,不乏此类错误。2017年全国图书馆联合编目中心工作会议上,国图联编中心就其数据质量所做的报告中提到,中心数据有4项问题还很突出。①数据硬伤。包括数据中错字、漏字,外文题名著录错误,字段、子字段、指示符错误,相关字段信息差异,以及尺寸、页数、装帧、书号、标点符号等著录错误。②选取图书题名的问题上没有统一认识、统一标准,不但直接影响图书内容的准确揭示,还会给用户查询带来不便,甚至误导读者。③对于同一套书或同一种书,如果选择不同的著录方式会产生大量重复数据,造成数据库混乱,浪费人力、物力,更严重的是误导使用者,给使用者带来困惑,这也是一直存在于联编数据库

中最凸显的问题。④数据库中已有的编目成果存在被盲目复制拷贝的现象。如果数据库中出现一条错误数据，随之会出现若干条同样的错误数据^[7]。除第3项外均为准确性问题。

另外，虽然图书馆界对CIP数据的准确性诟病较多，但又经常直接沿用CIP数据的分类标引结果，也会造成数据不准确。如《煤制天然气》的摘要中明确“本书是煤化工领域关于煤制天然气的学术著作”，CALIS数据未与编目实体核对，直接沿用了其CIP分类号F407.9（工业部门经济之建筑、水利工程），显然是不恰当的。

2.3 名称规范详略程度不同

根据世界书目控制有关原则，“国家书目机构有责任建立本国著者（个人和团体）的权威形式，并建立本国著者（个人和团体）的权威清单”^[8]。在我国，国家图书馆联合编目中心，CALIS联机合作编目中心，台湾地区的汉学研究中心，香港特别行政区的大学图书馆协作咨询委员会都开展了名称规范工作。国家图书馆和CALIS的规范库数据总量均超过130万条，台湾地区的汉学研究中心大约有70万条，香港特别行政区的大学图书馆协作咨询委员会超过20万条。但在规则层面差别较大，尤其在名称附加方面做法各异^[9]。以对“鲁迅”的规范为例（见表1），可以看到各家的详略程度不同。国家图书馆最简略；台湾地区的汉学研究中心则最详细，区分度更高。

名称规范库的建设和推广应用，对于用户检索效果影响很大，尤其是对检全率的影响更大。据笔者的不完全了解，高校图书馆在业务系统中使用名称规范的很少，尤其是一些地方类二本及以下层次的高校几乎完全没有使用。目前来看，名称规范库的普及推广还任重道远。国家图书馆与CALIS成员馆对其名称规范库的使用情况并不乐观，一方面是成员馆对数据的专业要求有待提高；另一方面，编目外包也对名称规范工作的正规化有影响。如何在编目外包时加强对数据加工商的名称规范要求，还需进一步探讨。

2.4 数据共享性差

（1）CIP数据格式与编目数据完全不共享。数据的共享性由数据的标准化程度决定，数据标准化程度越高，共享性越高。CIP数据、国图数据、CALIS数据采用不同格式，著录与标引项目也不同。传统CIP数据格式，用户只能查询浏览不能下载。2014年实施的CNOIX规定了图书产品信息的XML技术规则、格式规范和代码表等，著录与标引字段较传统CIP数据格式下要更全面。但不论是传统格式还是数字格式，CIP数据格式与下游环节目前采用的MARC数据格式完全不同，没有提供数据生产方与流通各相关方的数据交换渠道，图书产业链上下游共享数据的目标仍未达成。也就是图书流通环节的两种数据与CIP数据均不具备继承性。

（2）国图数据与CALIS数据间没有直接的数据交换与交流。流通环节的国图数据标准与CALIS数据标

表1 不同图书馆中文标目对比样例^[5]

	国家图书馆	CALIS (以简体中文为例)	台湾地区的汉学研究中心 (MARC21格式)	香港特别行政区的大学图书馆 协作咨询委员会
鲁迅	200#\$a鲁迅 \$f(1881-1936)	200#0\$7jt0yjt0y \$a鲁迅, \$f1881-1936	1001#\$d(清) \$a周树人, \$c中国文学	1001#\$aLu, Xun, \$d1881-1936
曹雪芹	200#\$c(清) \$a曹雪芹 \$f(1715/1724-1763/1764)	200#0\$7jt0yjt0y \$a曹雪芹, \$f约1715-约1763	1001#\$d(清) \$a曹雪芹	1001#\$aCao, Xueqin, \$dapproximately1717-1763
乾隆皇帝	200#\$c(清) \$a乾隆 \$f(1711-1799)	200#0\$7jt0yjt0y \$a清太宗, \$f(1711-1799)	1000#\$d(清) \$a清太宗	1000#\$aQianlong, \$cEmperor of China, \$d1711-1799
弘一法师	200#\$李叔同 \$f(1880-1942)	200#0\$7jt0yjt0y \$a弘一, \$f(1880-1942)	1001#\$a释弘一, \$c佛教, \$d1880-1942	1000#\$aHongyi, \$cda shi, \$d1880-1942

准都以ISBD为依据,且都有改进。国图标准更强调中国特色,CALIS标准更注重接近国际标准。不论这两种不同的倾向有怎样深厚的渊源,都造成依据它们生成的书目数据不利于共享与交换的现实局面。具体而言,二者在字段必备性上要求不同,在各字段的著录标引要求上也存在差异,这些造成流通环节书目数据的共享性差。

(3)数据的国际交换与共享问题。CIP数据标准与《普通图书著录规则》国家标准保持一致,不采用任何国际标准,基本没有国际共享的可通用性。新的XML标准有望实现国际共享,但其在国内出版界仍未推行,对于国际共享仍很遥远。编目数据按ISBD制定规则,其国际共享可能性更高。就目前来看,国图数据没有与国际合作的协议,CALIS有协议但数据交换并不频繁。

3 书目数据质量问题原因分析

3.1 各制作流程环节衔接不畅

图书全产业链上书目数据质量问题形成的主要原因,在于数据制作流程上各环节衔接不畅。CIP数据由版本图书馆制作,编目数据由图书流通环节的两大图书馆系统分别制定编目细则进行,且相互间没有正式的数据交换渠道。

3.2 缺乏质量严格把关

从数据制作流程可以看出,数据制作各环节都缺少严格的质量监控。

(1)CIP书目数据制作的缺点。生成于图书发行环节前,CIP数据制作受限于出版社提供的信息,这些信息既无作者参与,也没有专业的编目人员参与。版本图书馆编目人员只能根据这些信息给出CIP数据。既无核校流程,也无核校可能,更没有数据质量审核。CIP数据本身对数据的完整性要求低,因此数据的准确性与完整性均不能满足用户使用需求。

(2)编目数据制作的缺点。由于参与编目数据制作人员多,对图书内容理解的不同,以及各馆编目时“从旧从错”传统等因素影响,同一种书可以有多条编目数据。另外由于各联编中心对于同一种书的编目数据设置了不同的详略等级,所以同一种书可以有不同详略等级的多条数据,这就更增加了同一种书多条编目

数据的情况。几千个图书馆编制的上千万条书目记录,实际反映的中文图书只有100多万种^[5]。而国图数据与CALIS数据实行系统内共建共享,即由系统内有资质的编目员上传数据,然后由中心人员按不同情况进行抽查式审核而非全部审核。

4 制作流程视角下的数据质量提升对策

不同标准存在于不同书目数据的制作环节,造成社会资源浪费。当前在图书生产环节采用CIP标准制作书目数据,在流通环节又采用2个标准制作书目数据。而流通环节这2个标准,对于CIP数据的依赖性都不高,进一步造成3个标准各行其是的现状。CIP数据、国图数据、CALIS数据各自为政,浪费了大量的社会资源,数据质量参差不齐。对此,应该谋求将3个数据整合起来的途径,并协调数据建设与使用各方的利益,保障数据建设方权益,促进数据生产与使用的良性运行。

4.1 统一数据制作,保证完整级数据的唯一性

遵循一种图书应该只有一条权威且全面的标准书目数据的原则,应将现有的3个数据制作主体及其使用的标准进行整合,保证完整级数据的唯一性。

(1)平衡各方利益,推行书目数据共建共享。对于全部由版本图书馆完成数据制作中力量不足的实际问题,可以借鉴现行编目数据的共建共享思路,进行各方权益协调,实现数据建设的共建共享。①对数据实行有偿使用。具体来说,作为目前CIP数据制作的主力机构,成为一级数据中心,版本图书馆应角色转换为书目数据的管理机构。②在机构内建立数据建设联盟。联盟内按各馆特色,将书目数据制作任务进行分配,由专业的馆来做专业的书目数据。在制度上,可以增加出版社的呈缴本册数,其中一本归数据制作承担机构;在数据的使用上,对参与书目制作的机构有优惠使用规定。这样,才能既发挥现有编目人才优势,解决数据制作中的实际问题;又能通过联盟力量,保障数据质量;还可以通过有偿使用,保障数据良性使用。

(2)改进数据制作,使数据上下游一体化。国家新闻出版广电总局规定,现有CIP依据《普通图书著录规则》《文献叙词标引规则》《中国图书馆图书分类法》和《汉语主题词表》对图书进行著录、分类标引、

主题标引^[10]。实际工作中,由中国版本图书馆承担的数据制作在分类与主题标引上严格遵循这一要求,但由于“手中无书”,仅根据出版社提交的工作单制作数据,显然不能做到完整、准确。为保证图书付印时版权页数据准确,又保证使数据实现上下游一体化,出版应在提交工作单时请作者对内容提要项把关,提供该书的自由主题词,并在申请制作数据时提交预印本。这样数据制作人员“有书在手”,有作者对内容提要图书主题的认可,就可以对图书做更完整、准确的揭示。数据应尽可能做成完整级数据,方便下游使用者根据需要做减法而不是加法,这样可以最大程度地保证数据的标准化,有利于数据交换。

4.2 在出版社、CIP中心、书商、图书馆间构建有效的数据交流渠道

目前,CIP数据发布在中国版本图书馆主页能以CIP核字号查询,但传统CIP发布后只提供打印格式与书目详细2种浏览格式,不提供数据下载。其馆藏资源查询只能对上一年1月1日至当前日期的数据进行查询^[11]。2014年起实施的CNOIX标准,将传统CIP表达升级为XML格式,且项目更丰富,在打通上下游数据共享环节上迈出了一大步。

构建一级数据中心与各二级数据中心(指为行业资源交流与管理为主要职能的国图联合目录中心、CALIS联合目录中心)之间以及数据中心与书商的数据传送通道是经济高效的数据共享方式。数据中心将已发行图书的数据向国图联合目录中心、CALIS联合目录中心开放,各图书馆从各自所属的联合目录中心下载书目数据,进行个性化处理,导入OPAC系统,方便本馆资源管理与用户检索。各馆在下载数据的同时,提交本馆馆藏信息,即可在联合目录中心体现。这样,既保证了数据统一性,也不影响各联合目录中心的资源共享。数据中心同时将待发行图书书目信息向书商开放,书商可为用户提供更准确的订购依据。

国图联合目录中心、CALIS联合目录中心从一级数据中心下载数据的费用实行中心先行支付,再与成员馆约定结算的方式进行。书商下载数据直接与一级数据中心结算。一级数据中心将收到的数据下载费与出版社按约定结算。这样,出版发行部门、书目控制部门都能从数据中获益,有利于提高数据质量。

4.3 重视名称规范库建设,发挥名称规范库作用

从世界范围的实践来看,为全部著者建立规范形式的趋势并不乐观。这个状况的形成,一是理论上存在必要性争议,二是现实中存在可能性障碍。理论上的必要性争议主要是指对于所有著者建立规范名称的现实意义是否足够大;现实中存在的可能性障碍是区分难度与成本太大,而使用价值太小。

从源头统一制作书目数据、构建现有3个数据中心及其他各相关方数据通道实现的基础上,名称规范工作的现实意义与使用价值得到了增强。统一规则,通过共建共享的机制建设名称规范库,并将名称规范文档与书目文档建立连接,形成规范系统,将进一步提高标准化书目数据的检索效率。在统一规则下整合现有规范成果是最可行的方法。名称规范库建设中,可通过利用维基百科条目的丰富信息来补充完善当前中文名称规范文档附加信息不足、名称难以准确区分的问题^[11],可通过实现“海峡两岸暨香港”在中文名称规范方面的合作^[12],谋求中文名称规范文档与虚拟国际规范文档的共享^[13]途径,逐步完善名称规范库。

参考文献

- [1] 王洋. 全国图书馆联合编目中心2017—2018年度工作报告 [EB/OL]. [2018-09-23]. <http://olcc.nlc.cn/index.php/page/document.html>.
- [2] 全国图书馆联合编目中心 [EB/OL]. [2018-05-03]. <http://olcc.nlc.cn/index.php/page/document.html>.
- [3] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会. 中国出版物在线信息交换 图书产品信息格式: GB/T 30330—2013 [S]. 北京: 中国标准出版社, 2014.
- [4] CALIS中文图书编目业务培训教材 [Z]. 北京: CALIS联机合作编目中心, 2011.
- [5] 陈源燕. 中文图书ECIP与自动编目手册 [M]. 北京: 书目文献出版社, 2003: 26.
- [6] 关于继续加强CIP工作, 进一步提高数据质量的函 [EB/OL]. (2011-04-02) [2018-10-01]. <https://www.capub.cn/zbbm/xgwj/04/1124.shtml>.
- [7] 刘小玲. 数据质量分析报告 [EB/OL]. [2018-06-21]. <http://olcc.nlc.cn/index.php/page/document.html>.
- [8] ANDERSON D. Universal Bibliographic Control; A Long Term

- Policy, a Plan for Action [M]. Pullach/MUnchen: Ver-lag Dokumentation, 1974: 47.
- [9] 王彦桥, 王广平. 中文名称规范数据的维护与整合 [J]. 图书馆杂志, 2017, 37 (2): 56-59.
- [10] 为什么要指定CIP数据的专门制作单位? [EB/OL]. [2018-04-29]. <https://www.capub.cn/zbbm/ywjd/04/1126.shtml>.
- [11] 中国版本图书馆(中央宣传部出版物数据中心) [EB/OL]. [2018-04-29]. <https://www.capub.cn/sjjs/index.shtml>.
- [12] 贾君枝, 薛秋红. 中文人名名称规范档与维基百科的链接 [J]. 图书情报工作, 2015, 59 (16): 129-134.
- [13] 贾君枝, 石燕青. 中文名称规范文档与虚拟国际规范文档的共享问题研究 [J]. 中国图书馆学报, 2014 (6): 83-92.

作者简介

原小玲, 女, 1972年生, 硕士研究生, 副研究馆员, 研究方向: 信息组织与社会信息化, E-mail: TYYUAN9333@163.com.

Bibliographic Data Analysis from the Perspective of Data Production

YUAN XiaoLing

(Library of Taiyuan Science and Technology University, Taiyuan 030024, China)

Abstract: There are problems in the data completeness, accuracy and accessibility of Chinese book bibliographic data. The main reason for this is that different standards are used by multiple agencies. To establish a nationwide unified bibliographic data based on CIP, to establish data channels between publishing houses and downstream, and to pay attention to the construction and promotion of the name specification base, can not only make easier for data exchange and sharing, but also can greatly save social costs. It can also truly reflect the relationship between the volume of publications and the library's collections.

Keywords: Chinese Book; Bibliographic Data; Data Quality

(收稿日期: 2018-10-14)

会议报道

第二届全国信息与文献标准化研讨会 在北京召开

2018年11月16日, 由全国信息与文献标准化技术委员会、全国图书馆标准化技术委员会联合举办的“第二届全国信息与文献标准化研讨会”在北京召开。来自全国图书馆、情报机构、档案馆、出版机构、博物馆、信息中心等信息与文献领域工作者共180余人参加研讨会。全国信息与文献标准化技术委员会秘书长曾建勋、全国图书馆标准化技术委员会秘书长申晓娟共同主持会议。

研讨会邀请南京大学信息管理学院苏新宁教授、北京大学图书馆肖珑教授、中山大学资讯管理学院肖鹏副研究员、上海图书馆张轶馆员分别就学术资源库建设与评价、中美电子资源国家标准比较研究、欧美公共文化机构年报编制、馆藏资源文献保存与保护作了标准化主题发言。会议就《科技人才元数据元素集》《引文数据库数据加工规则》《图书馆及相关组织的注册服务》《CNMARC的XML表示》《中国机读馆藏格式》《中国机读书目格式》《公共图书馆少年儿童服务规范》《图书馆视障人士服务规范》《流动图书车车载装置通用技术条件》《图书冷冻杀虫技术规程》《古籍函套技术要求》《图书馆数字资源长期保存信息包封装规范》等12项信息与文献国家和行业标准集中进行了宣贯。

全国信息与文献标准化技术委员会每两年召开一次标准化研讨会, 对扩大信息与文献标准化影响力, 促进信息与文献标准的实践应用发挥了积极作用。