中文社会科学引文索引数据质量问题研究

张友谊 刘 春

(重庆师范大学图书馆 重庆 400047)

摘 要 中文社会科学引文索引(CSSCI)查询和统计功能的有效实现,是以数据的准确、规范为基础,笔者从数据结构和数据内容两个角度详细分析CSSCI的数据质量问题,并提出了改进建议,以期CSSCI进一步发展完善。

关键词 中文社会科学引文索引 数据质量 改进措施

中图分类号 G353.21 **文献标识码** A **文章编号** 1002-1965(2012)01-0021-04

Research on the Data Quality Problems of CSSCI

ZHANG Youyi LIU Chun

(Library of Chongqing Normal University, Chongqing 400047)

Abstract Realization of the query and statistical functions in Chinese Social Science Citation Index (CSSCI) is based on the accurate and standard data. From the data structure and data content, the authors analyze data quality issues of CSSCI and propose the corresponding measures for improvement, with a view to the further development of CSSCI.

Key words chinese social science citation Index data quality improvements

中文社会科学引文索引(Chinese Social Sciences Citation Index,以下简称CSSCI)是由南京大学中国社会科学研究评价中心开发研制的引文数据库,用来检索中文人文社会科学领域的论文收录和被引用情况[1]。作为我国人文社会科学主要文献信息查询的重要工具,填补了我国社会科学成果量化统计和评价工具的空白,具有重要的科学价值和实际应用价值。CSSCI查询和统计功能的有效实现,是以结构合理、著录准确规范的数据为基础,而笔者最近在利用CSSCI统计分析数字图书馆主题文献的过程中发现,来自CSSCI的数据存在诸多质量问题,这些问题可大致分为结构和内容两个方面。

1 CSSCI数据结构存在的问题

1.1 年份数据结构问题 很多数据挖掘工作不能直接在CSSCI检索主界面实现,研究人员常需将检出记录详情下载后按字段类别进行拆分、连接、归并等处理,以达到特定统计分析目的。但笔者对比1998-2010的来源文献库下载数据结构,发现各年代库提供的记录结构差异极大,字段个数不等,名称不统一,甚至个别字段名称都弄错。1998-2010各年代来源库的检索结果下载数据结构详见表1。

从表1可看出,不同年代来源库下载记录结构存在以下问题:

a. 不同年代记录的字段数不同。2009提供的记录多达19个字段,其他年代字段数在15-18间无规律的变动,记录结构不统一导致的结果是研究者不能把各个年代下载数据合并在一起统一进行格式转换处理,如果有多年数据要处理,只能先按年代把相同年份的下载数据(txt文档)合并在一起单独进行格式转换,然后分别从每年的数据中抽取所需字段,导致工作效率大大降低。

b. 字段名称不统一。比如同是著录来源文献篇名的字段,在2002-2004和2008中的字段名是“来源文献”,其他年份则叫“来源篇名”;同是存放文章分类号的字段,在1999和2010中的字段名是“期刊代码”,其他年份叫“中图类号”,而我们平常理解的“期刊代码”是指期刊的CN号或ISSN号,2009年数据竟把该字段名误为“中途类号”,显然是录入字段名时用拼音输入法产生的失误。在同一检索系统中,相同字段相同名称应是最基本的规范要求。

c. 字段名称指代不清。比如“标引词”这一字段

收稿日期:2011-08-06 修回日期:2011-12-16

作者简介:张友谊(1978-),女,硕士,馆员,研究方向:信息咨询;刘 春(1972-),男,硕士,馆员,研究方向:网络信息检索。

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表1 来源库各年代下载记录的字段构成

年代	记录的字段构成	字段数
1998	文件序号、来源篇名、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、中图类号、年代卷期、标引词、参考文献	15
1999	文件序号、来源篇名、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、期刊代码、年代卷期、标引词、参考文献	15
2000	文件序号、来源篇名、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、地区、学科分类、第一作者、中图类号、年代卷期、标引词、参考文献	16
2001	文件序号、来源篇名、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、中图类号、年代卷期、标引词、参考文献	15
2002	文件序号、来源文献、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、中图类号、年代卷期、标引词、参考文献	15
2003	文件序号、来源文献、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、中图类号、年代卷期、标引词、参考文献	15
2004	文件序号、来源文献、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、中图类号、年代卷期、标引词、参考文献	15
2005	文件序号、来源篇名、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、地区、学科分类、第一作者、中图类号、年代卷期、标引词、学位分类、基金类别、参考文献	18
2006	文件序号、来源篇名、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、地区、学科分类、第一作者、中图类号、年代卷期、标引词、基金类别、参考文献	17
2007	文件序号、来源篇名、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、中图类号、年代卷期、标引词、参考文献	15
2008	文件序号、来源文献、英文篇名、来源作者、作者姓名拼音、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、中图类号、年代卷期、标引词、参考文献	16
2009	文件序号、来源篇名、英文篇名、来源作者、文章类型、基金、期刊、第一机构、机构名称、地区、学科分类、第一作者、中途类号、年代卷期、标引词、标志、学位分类、基金类别、参考文献	19
2010	文件序号、来源篇名、英文篇名、来源作者、作者姓名拼音、文章类型、基金、期刊、第一机构、机构名称、学科分类、第一作者、期刊代码、年代卷期、标引词、基金类别、参考文献	17

名，常说的“标引词”是指用来标识文献内容的确切的检索标识，主要包括分类号、标题词、叙词、关键词、人名、地名等，而表中的“标引词”实际指的是论文关键词，建议该字段名直接用“关键词”；“地区”字段内容对应的是地区行政代码，并非地区名称，建议该字段名称改为“行政地区代码”；“年代卷期”字段内容还包括了文章在期刊的页码信息，应改为“年代卷期页码”，使其指代更准确。2009年的下载数据中增加的“标志”字段也指代不清，笔者经多方考证，均不知该“标志”指的什么标志，也不知它对信息检索或科研评价有什么用处？

1.2 “显示”和“下载”记录格式问题 在来源文献的检索结果界面，使用者可以通过点击“显示”和“下载”两种方式获取选中检索结果的详细信息，点击“显示”将会在新开的窗口中罗列选中记录的12个字段内容（详见表2），每个年代记录点击“显示”都展示相同结构数据；而点击“下载”得到TXT文件，不同年度TXT文件中记录的字段数各不相同（如表1所示），并且字段名与“显示”结果也不一致。表2以点击“显示”所展示的字段为基础，罗列出“下载”数据中的对应字段名称。CSSCI来源库在“显示”和“下载”两处展示相同记录详情时所用字段名差异较大，统计的12个字段就有8个字段所用名称不一致，常让使用者无所适从。

1.3 检索入口名称问题 CSSCI的来源文献库和被引文献库的检索界面提供多个检索入口，每个独立

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2　“显示”和“下载”记录结构对比

“显示”记录字段名	篇名	英文篇名	作者	作者机构	文献类型	学科类别	中图类号	基金项目	来源期刊	年代卷期	关键词	参考文献
对应“下载”记录字段名	来源篇名	英文篇名	来源作者	机构名称	文章类型	学科分类	中图类号	基金	期刊	年代卷期	标引词	参考文献
名称是否一致	否	是	否	否	否	否	是	否	否	是	否	是

表 3　检索入口名称与对应下载数据的字段名对比

来源检索入口名称	学科类别	学位分类	文献类型	作者	作者机构	作者地区	期刊名称	年代卷期	基金类别	基金细节
对应下载数据字段名	学科分类	学位分类	文章类型	来源作者	机构名称	地区	期刊	年代卷期	基金类别	基金
名称是否一致	否	是	否	否	否	否	否	是	是	否
引文检索入口名称	被引文献作者	被引文献期刊	被引文献类型	被引文献篇名(词)	被引文献年代	被引文献细节				
对应下载数据字段名	被引文献作者	被引文献期刊	文献类型	被引文献篇名	被引期刊年代	被引文献出处				
名称是否一致	是	是	否	否	否	否				

检索入口对应的是数据库记录中一个特定字段。而我们下载的检索结果记录也是由数据库中对应文献记录的多个字段组成。如果两处都对应的是相同字段,它们所用字段名称也应该相同,才便于理解使用,但实际情况并非如此,表 3 把检索入口名称和下载数据中的对应字段名称列表对比。

从表 3 可知,10 个来源检索入口名称就有 7 个与下载数据的对应字段名称不统一,虽然他们指向的是同一字段内容,但表达形式各异。CSSCI 被引文献检索界面提供 6 个检索入口,其中有 4 个入口名称与对应下载数据字段名称不统一,其中检索入口“被引文献细节”字段名指代不清,对应的下载数据字段名“被引文献出处”表达也不准确。令人欣慰的是,CSSCI 在检索结果页面提供二次检索入口,可以对初次检出结果进行再次检索,二次检索入口的字段名与下载数据的字段名完全一致,系统在这个两个地方的对应工作做得很好。

2　CSSCI 数据内容存在的问题

CSSCI 数据结构方面的问题,可以通过增加或减少字段、规范字段名称等方式统一调整,修改校对工作量相对不大。而最影响 CSSCI 使用效果,问题最多的是 CSSCI 各个数据库中记录字段内容的差错,它们将直接影响到查询和统计评价结果。

2.1　公共代码字典库的数据错误　CSSCI 设计公共代码字典主要用来存放系统中使用到的一些代码型数据。这些数据类型多,均自成一体,如,地区代码、机构类别、文章类型等[2]。笔者利用 CSSCI 统计重庆地区 2001–2004 的发文量,在“作者地区”检索入口输入“重庆”进行检索,找到结果为 0(显然与事实不符),系统返回的检索结果提示为“没有符合您的检索记录! 请重新检索! LY01,LY02,LY03,LY04,: DQ = 5102 $”,从提示中的“DQ = 5102”可以分析出 CSSCI 的地区代码字典库把重庆的地区代码映射为 5102; 笔者返回检索界面,把检索词“重庆”换成“四川”,点击检索,系统返回的检索结果提示为“命中结果 8282 篇,检索表达式: LY01,LY02,LY03,LY04,: DQ = 51 $”,同样分析可知,CSSCI 的地区代码库中“四川”对应的代码是 51; 对比两地的地区代码可知,在 CSSCI 的地区代码表中,重庆还是四川下面的一个二级城市,而重庆 1997 年就已经成为直辖市,笔者再查询国家统计局的行政区划代码表,重庆是 50 万,四川是 51 万,这说明 CSSCI 的地区代码库用的是老古董数据,并且一直未对该库数据进行更新维护,这样会导致 CSSCI 的地区发文量统计排名严重失真。由于时间精力限制,笔者并未对其他行政地区代码和其他代码字典库进行核对,不知是否也存在类似问题。

2.2　来源库的字段内容问题　来源库记录全为期刊论文,字段内容相对比较规范,不过还有以下问题需要注意。

a. 部分年代的“文章类型”字段数据为空。笔者对 1998–2010 的下载数据进行逐年分析,发现 2001–2004 的下载数据中“文章类型”字段为空,即系统未著录该字段内容。这样导致的结果就是没法分别统计 2001–2004 年收录的来源文献中“论文、综述、评论、传记资料、报告”的数量和比例,如果检索这 4 年的数据时限定了具体的“文献类型”,找到的结果将为 0,导致相关统计结果严重失真。

b. 来源库的“机构名称”字段著录不规范。该字段完整规范的著录应是“机构名称+部门名称,邮政编码”,如有合著机构,则用斜杠“/”连接,类似“承德医学院图书馆,067000/桂林理工大学图书馆,541004”,其中的“承德医学院”为机构名称,“图书馆”是部门名称。但在来源库中“机构名称”字段著录形

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

式多样，有著录齐全的，有只著录机构名的，有只著录机构和部门名，有只著录机构和邮编的，很不利于研究人员对机构和部门数据的准确抽取和统计分析。建议把“机构名称”拆分为三个字段，分别是“机构名称”、“部门名称”、“邮政编码”，或在“机构名称”字段下面设置对应的三个子字段，若某子字段无内容，则用特定符号占位，以便于数据的规范录入和准确提取。机构名称字段的严格规范既有利于机构发文的准确统计，同时也有助于同名作者的区分，把作者名和机构名连在一起进行分类汇总，基本上就可以区分出同名作者情况，所以说机构名的规范是非常重要的。

除此之外，在来源库中还存在不少记录的“文献出处”、“学科分类”、“英文篇名”等字段出现著录不准确、不规范甚至未著录等问题，这些问题都会在一定程度上影响CSSCI的查询和统计结果。

2.3 引文库的字段内容问题 引文数据库的记录结构虽比来源库简单得多，但由于我国社会科学研究领域中的学术规范性和社科期刊的标准化水平不高，导致社会科学期刊的参考文献著录很不规范[3]，再加上CSSCI数据录入的差错，进一步导致CSSCI引文库字段内容差错百出。

a. 字段内容著录格式不统一。比如“被引文献出处”字段，当检索张晓娟的文章《论数字图书馆》，发现原本同一来源的文章，“被引文献出处”却有3种不同标注，1998年的引文库中标为“(1):3，1996”、1999年的标为“1996(1):2—7”、2001-2004的均标为“1996(1)”。笔者追溯各来源文献中对该引文的著录，均标注为“1996(1):2—7”，说明该问题的出现并非由于原刊标注不规范引起，而是CSSCI著录该字段的标准频繁变化或数据录入人员未严格执行著录标准造成的。

b. 字段内容的文字和标点符号录入差错。笔者在利用CSSCI进行某主题高频被引文献统计分析的过程中发现，原本同一篇文献，因为文献标题、作者、年代等字段著录的差错，在计算单篇文献累计被引频次时却被识别成多篇不同的文献，导致该文章的累计被引频次大大降低。“被引文献篇名”字段最长，该字段出错的情况也最多，常见错误包括文献标题出现多字、少字、错字、文字顺序排错、增加或减少标点符号、中英文标点颠倒、半全角符号数字颠倒、空格增减等。表4列出CSSCI引文库中找到的标题字段的一些常见错误，经核实表中的每组文献均指代的是同一篇被引文献，其中第一篇的内容表述是准确的。

表4仅列举了标题字段内容的部分出错情况，同样，作者字段中错漏也不少，比如同一作者“姚晓霞”有著录成“姚晚霞”的，“肖珑”有著录成“肖笼”的，译文文献中有的记录只著录译者，有的译者和原作者均著录，有的著录原作者中文译名，有的著录原作者外文名，而原作者外文名又出现不同的缩写方式，中文翻译名也不统一等。“被引文献出处”字段的年代、卷期、页码数据也有出错情况，剩下的其他几个字段出错情况相对较少，刊名字段最规范。

表4 被引文献标题字段出错情况例举

被引文献标题字段	出错情况
“数字化陷阱”与数字图书馆资源建设的可持续发展 “数字化陷阱”数字图书馆资源建设的可持续发展	第二篇漏掉“与”字
21世纪图书馆发展的方向——数字图书馆21世纪图书馆发展的趋势——数字图书馆	“方向”误录为“趋势”
IBM数字图书馆中图像的处理与加工IBM数字图书馆中的图像处理与加工	第二篇“的”字位置不对
J2EE和.NET的比较——构建数字图书馆系统平台的思考J2EE和NET的比较——构建数字图书馆系统平台的思考	第二篇漏“.”
数字图书馆——原理与技术实现数字图书馆:原理与技术实现数字图书馆原理与技术实现	第二篇把“——”误作“:”，第三篇漏掉“——”
21世纪数字图书馆联盟:香港JULAC(大学图书馆长联席会)实例21世纪数字图书馆联盟:香港JULAC(大学图书馆长联席会)实例	第二篇把半角“()”录为全角

3 改进建议

CSSCI查询和统计分析功能的有效实现，数据库记录结构合理，字段内容著录准确、规范是基本前提，为此，笔者提出以下改进建议。

第一，完善建库工作的标准和规范，并在工作中遵循统一的标准、规范；加强数据录入和校对人员的培训，强化责任意识和质量意识。上面提到的许多问题并非技术难题，而是意识是否到位，责任是否落实，有无规矩可依的问题。

第二，统一“检索入口”、“下载”数据、“显示”数据及“二次检索入口”的字段名称，规范每个字段的具体名称，使其指代清楚，表意准确，如“标引词”改为“关键词”，“期刊代码”改为“中图类号”，“地区”改为“行政区划代码”等。同时，在CSSCI主页显著位置提供帮助链接，对系统各个字段含义进行说明，帮助使用者了解数据结构。统一各年代来源库下载记录的字段个数，如果新的数据要增加字段，以前年代的数据至少也要先增加一个相同的空字段，以保持数据结构的一致，并尽快回溯补充相应字段内容。

第三，定期维护和更新代公共代码字典库数据，避免文中提到的类似错误。

第四，题名和作者等难以建立字典库的字段严格

（下转第46页）

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.4 只有线性评价值反映了“可实现的极大值目标” 在线性评价结果中,满分100是一种可实现的极大值目标。如果评价结果极大值是100,说明排名第一的评价对象各项指标都排在第一,达到最优值;如果评价结果极大值不是100,说明排名第一的评价对象部分评价指标排在第一,部分评价指标没有排在第一,但是有其它评价对象这些指标排在第一。总体上可以说,所有评价指标的极大值都有评价对象曾经达到过,或者说,排名第一的评价对象得分与满分100的差距相当于暂时的绝对差距,这种差距客观上是可以通过努力达到的,不是一种不可能实现的目标,因此满分100是一种可实现的极大值目标。

对于非线性评价结果而言,最好的评价对象都标准化为满分100,因此评价结果并没有反映出某种绝对目标,即使对最好的评价对象也是如此。因此,从这个角度讲,线性评价结果比非线性评价结果提供的信息要更为充分。

参 考 文 献

[1] Economic and Social Commission for Western Asia, United Nations. New Indicators for Science, Technology and Innovation in the Knowledge-based Society [R]. New York,2003

[2] National Science Foundation. Science and Engineering Indicators [R]. Washington, 2002

[3] http://www.rae.ac.uk

[4] 李晓轩. 德国科研机构的评价实践与启示 [J]. 中国科学院院刊,2004,19(4):274-304

[5] International Institute for Management Development. World Competitiveness Yearbook [R]. Lausanne 2004

[6] OECD. Main Science and Technology Indicators [S]. Paris,2003

[7] 王 俭. 科研机构专家评价方法研究 [J]. 科研管理,2005(4):152-155

[8] 关忠诚,赵二帅. 偏好DEA在科研机构相对效率评价中的应用 [J]. 科学学与科学技术管理,2008(8):40-45

[9] 李 强,韩伯棠. 建立新公共科研机构绩效评价测度体系研究 [J]. 科学学研究,2006(4):243-248

[10] 孙 俐,苏晨光. 多目标规划在科研机构管理评价的应用 [J]. 天津理工学院学报,2004(9):72-74

[11] 赵红专,翟立新,李 强. 公共科研机构绩效评价的指标与方法 [J]. 科学学研究,2006(2):85-90

[12] 孟 薇,李 强,刘文斌. 基于3E理论构建科研机构评价指标体系 [J]. 科学学研究,2007(10):908-914

[13] 张伟倩,缪 园. 组合评价模型在我国国立科研机构绩效评价中的应用 [J]. 科学学与科学技术管理,2008(4):36-40

[14] 俞立平,潘云涛,武夷山. 学术期刊综合评价数据标准化方法研究 [J]. 图书情报工作,2009(6):136-139

[15] 李美娟,陈国宏,等. 综合评价中指标标准化方法研究 [J]. 中国管理科学,2004(10):45-48

[16] 刘东波,陈玉娟,等. 灰色评价指标的一种曲率可调非线性标准化转换方法 [J]. 华东理工大学学报:自然科学版,2009(2)

(责编:王平军)

(上接第24页)

采取双轨录入制度,在双轨录入的基础上再进行软件和人工纠错。从目前统计情况来看,CSSCI现行采取的单轨输入加数据校对机制还未能把数据出错率保持在较低水平,数据质量无小事,并非是“许多差异对于数据的质量是无关紧要的(如符号、数字和空格的半角与角,西文的大小写等)”[4]。在数据库中可能就因为一个小的符号差异导致原本该是相同文献却被系统识别成不同文献,原本该是命中的检索结果而实际却未被检出,影响查询和统计评价结果客观准确。

4 结 语

如果数据质量出现问题,将直接影响到检索项的查全率和查准率,从而造成统计数据偏差,严重影响评价结果的权威性和客观性。为保证数据质量,CSSCI研制中心各个环节的工作人员都要有强烈的责任意识,把好结构设计和数据录入的各道关口。但很多数据质量问题是由于原作者和原期刊标注错误造成的,所以单靠CSSCI研制中心一家是难以圆满解决问题,还需国家新闻出版署、各学科学会、期刊研究会、广大社科研究人员以及期刊编辑人员都采取措施,增强标准化和规范化意识,严格执行《文后参考文献著录规则》(GB7714-87)和国际文献著录规则,严格按要求列出参考文献及其各项著录信息,尽快改善和提高我国社科期刊的标准化和规范化水平[5]。同时,CSSCI在调整来源刊时可考虑加重参考文献著录规范程度的考察权重,以督促期刊标注规范化。相比国外的SCI和SSCI,CSSCI还算是一个新生事物,本文不是要对它求全责备,而是想提出问题引起大家对它的关注,促进它不断发展完善。

参 考 文 献

[1] 中文社会科学引文索引(CSSCI)简介 [EB]. 2011-08-04. http://219.219.114.10/news_show.asp?Articleid=119

[2] 苏新宁. 中国社会科学引文索引设计 [J]. 情报学报, 2000(4):290-295

[3] 李志鹏. 中文社会科学引文索引的不足与改进 [J]. 情报资料工作, 2005(5):108-111

[4] 苏新宁. 引文索引数据质量控制研究 [J]. 中国图书馆学报, 2001(2):76-78

[5] 邱均平. 中文社会科学引文索引的意义、特点与改进建议 [J]. 武汉大学学报(社会科学版), 2002(2):232-236

(责编:贺小利)

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net