地方综合年鉴数据处理探析

——以《梧州年鉴》为例

赖红柳*

摘　要　年鉴作为严谨科学的资料性著述,数据的准确性至关重要。地方综合年鉴的数据反映了地情状况。由于年鉴数据来源的多样性,常会导致数据的相互矛盾和不实以及使用不规范等问题,对数据进行技术处理,即采取验算、核算和逻辑核对等步骤必不可少。在数据技术处理的具体方式、方法上,注重数据使用的规范以及表达的完整,确保数据处理关联性和完整性。由于年鉴数据的真实性和重要性特点,提高数据处理人员的专业化水平迫在眉睫。

关键词　地方综合年鉴　数据处理　大数据

地方综合年鉴的数据,是从数量的角度来记述一个行政区域内自然、政治、经济、文化和社会情状的资料,是年鉴中不可或缺的一部分。在年鉴的经济类目中,数据的地位尤为举足轻重。大到国民经济的主要成就,小到某项工程的进展情况,往往都用数据来表述。如果数据上出现问题,年鉴质量就不能保证,影响其权威性。年鉴质量的好坏很大程度上取决于数据处理技术水平的高低,取决于"数据本身的准确性、使用的规范性以及表达的完整性"①。为使年鉴成为严谨科学的资料性著述,必须把好数据关。笔者结合《梧州年鉴》编纂出版的工作实践,就如何处理地方综合年鉴的数据提出自己的一点浅见。

一、数据来源的多样性是年鉴客观存在的现实

出版地方综合年鉴是一个系统地、连续地、需要各个部门、单位配合的政府工程,是政府行为,所谓众手成书。因此,年鉴数据的来源渠道也是多样的。可分为传统来源和创新

* 赖红柳,女,广西壮族自治区梧州市人,梧州市地方志编纂委员会办公室编辑出版科副科长、编辑,主要研究方向为志鉴、地方史。

① 杨汉成:《统计数据在年鉴中的规范化运用》,《山西统计》2003 年第 11 期。

来源两种。考察年鉴的数据来源,有助于我们理解数据处理的因由。而且,在具体数据处理上,能够根据数据的不同来源做出正确的判断。

(一)传统来源

从年鉴资料来源看,一般地方综合年鉴的数据来源主要由党政部门、企事业单位、社会团体及行业协会等按年鉴编纂要求提供,多数来自会议讲话、会议纪要、工作报告、工作总结、统计报表、工作台帐、新闻通稿等。此外也有一部分数据资料由年鉴编辑人员"通过新闻报刊、查阅档案和提炼网络信息以及调查访问等方式进行搜集"①。这往往因为数出多门、口径的不同很容易导致数据出现矛盾。权威的数据来源是统计部门公布的数据。实际上,统计部门每年下半年9月至10月才公布所有上一年的数据,《梧州年鉴》在统计部门尚未公布数据之前,年鉴中的重要数据全部空缺,等到统计部门公布后再填补。如此,难以做到在出版年度的上半年内出版。

(二)创新来源

"当前年鉴资料的来源绝大多数还是以单位供稿为主导,被动收集资料为主,网络信息来源利用很少,内容也主要是国家机关治理涉及的领域,在政府部门职能不涉及的一些社会和行业领域,内容则很少。"②2015年8月31日,《国务院关于印发促进大数据发展的行动纲要的通知》出台。随着我国大数据发展规划的落实,"云计算"技术的广泛应用,年鉴应逐步突破单一被动接收资料的模式,把一些党政机关职能范围外,以前难以出现在年鉴中的一些内容,"如在一些服务业领域像电子商务发展情况、餐饮行业客流量等数据,均需依托相关行业数据库来收集整理"③,弥补年鉴记录的空白。

由于年鉴数据来源的多样性,数出多门,常常导致数据的矛盾和不实,以及使用的不规范等问题。因此,有必要对数据进行技术处理。

二、数据的准确性是保证年鉴客观记述的根本

数据资料用以反映事物发展变化的程度。数据的准确程度,直接关系到年鉴的质量和使用价值。在编纂年鉴过程中,因为数出多门,也由于各种主客观原因,提供的数据不够准确,甚至有虚假成份。因此要确保数据的准确性,必须对数据进行多方核实。

(一)验算核实

我们核实数据是否准确,常常需要根据各种公式进行演算。因此,年鉴工作者需要掌握一些统计学、经济学方面的常识。例如,人口统计方面的自然增长率、平均期望寿命,运输量的人/公里和吨/公里等都是用特定的公式计算的特定概念,只有弄清它们的概念和计算方法,才能进行正确的演算核验。

---

① 《地方综合年鉴编纂出版规定》(中指组字〔2017〕6号),2017年12月21日。

② 张军:《大数据时代年鉴面临的挑战与机遇》,《江苏地方志》2016年第2期。

③ 张军:《大数据时代年鉴面临的挑战与机遇》,《江苏地方志》2016年第2期。

对存在包含关系的一组数据或图表,如果总项和分项都罗列完全,就需要计算核验单项相加是否与合计相等。例如,“全市有党员 14.09 万名。其中,党政机关工作人员 1.23 万名,企事业单位、民办非企业单位专业技术人员和管理人员 2.59 万名,工勤技能人员 2963 名,农牧渔民 6.13 万名,学生 716 名,离退休人员 2.60 万名,其他职业人员 7066 名”。单项相加与合计不符,经查其总数是错误的。

关于绝对数与比率的核算,尤其要注意有些概念相近但是计算方法不同、特别容易混淆的情况。比如,“百分比”不等于“百分点”,“增长了”不等于“增长到”,相当于多少倍不等于增加了多少倍,增长两倍不等于翻两番,递增率不能用简单算术平均法计算而要用几何平均法计算,等等。

(二)对照核实

同一内容、同一时间内的数据,尤其是涉及国民经济和社会发展的基础数据,要与统计部门及相关主管部门的数据对照核实。《地方综合年鉴编纂出版规定》明确要求:“年鉴采用的数据应以统计部门提供的为准,未列入统计范围的,以业务主管部门的为准。数据不一致时,应加以说明。”

年鉴经常有同一数据在全书多次出现的情况,概览、概述中的数据要与各类目、分述中的数据相对照,不同类目出现的同一项数据也应对照核实。发现有不一致的,应请供稿人员重新查对。在资料来源、统计范围、统计口径、计算方法及行业分类等方面找出原因,务求一致,使数据更具可比性和说服力。在编纂过程中,应注意文与表、文与图、图与表之间数据的一致性。内文与图表的数据相对照,不一致的地方都应该核实纠正。

年鉴中同一数据不同地方出现不一致的情况比较普遍,即使统计部门出的统计年鉴也不例外。比如《梧州统计年鉴(2016)》“概述”中的梧州市行政区域总面积是 12588 平方公里,而在其后的分述表格中,市总面积是 12572.48 平方公里。通过查问得知,原来是因为市国土局重新核定了全市及各县(市)区的行政区域面积,而《梧州统计年鉴(2016)》在分述表格中采用的是新数据,但是概述中仍沿用旧资料,所以导致出错。

(三)逻辑上核实

除以上两种核实外,相关数据还应根据自己的见识,从逻辑角度进行检查,看看各项相关数据的大小或变化是否符合逻辑。如果数据过大或过小,不合常理,就有可能是错误的。例如,2016 年岑溪古典鸡养殖,年饲养量 1500 多万羽,年出栏 2000 多万羽。计算公式是“年饲养量 = 出栏量 + 存栏量”,年饲养量不可能比年出栏量少,所以明显有误。经供稿单位核查,数据统计时把非古典鸡的年出栏数也算进古典鸡的年出栏量里从而导致错误。

在行政区划发生变化的情况下,虽然原来的名称不变,但区划变化前后的数据不能混同使用,也没有可比性。比如 2013 年梧州市调整部分行政区划,撤销万秀区、蝶山区,设立新的万秀区;苍梧县分出龙圩区。在《梧州年鉴(2014)》中,苍梧县和新万秀区的有关全县(区)的数据,再与上年数据进行对比分析或相加等,都是不科学的,应注意删除修改。

三、数据使用的规范化

由于数据来源的多样性,年鉴收集来的原始数据的表现形式也五花八门、杂乱无章,这就需要我们对年鉴中数据的书写形式和计量单位用统一标准规范起来,以消除混乱,方便读者使用。

(一)数字书写的规范是年鉴规范化的基础

为保证年鉴数据使用的规范化,年鉴的数字书写应当按照《出版物上数字用法》(GB/T 15835—2011)执行。这个规定适用于除文艺类和重排古籍外的所有出版物。但查阅各地年鉴,发现数字书写仍然存在许多不规范的地方,各地标准不一,尤其是对多位数的处理。

1. 多位数的处理

对照上述《出版物上数字用法》,各地年鉴比较容易忽略的规定是,整数部分每三位一组,以“,”或组间空 1/4 个汉字分节,小数部分不分节。四位数以内的整数可以不分节,但所查年鉴完全不分节的居多。对于多位数的处理,应倾向于沿用广西第一轮修志行文规范中的做法,除图表外的内文数字,超过万的数字以万为单位,超过亿以亿为单位,同时小数点后以四舍五入方式保留两位数字。原因有四:

一是年鉴收录数据的比重较大,一部年鉴中的数据资料约占全书的 30%,尤其是经济部类所占的比重更大,某些类目全由数据说明问题。比如“财税金融”中的“财政收支”类目。如果全是不作处理的多位数,显得冗长繁琐。

二是年鉴收录的原数据五花八门,有以个位为单位的多位数,有以万为单位的,有以亿为单位的数;有保留到小数点后一位数的,同时也有保留两位、三位甚至四位数的,杂乱无章。

三是按四舍五入保留小数点后两位得到的数字,在年鉴常见的分项合计运算中,和处理前的精确数相比较,纵使有误差也在处理后数字的 0.01 范围内,也就是说误差不过 0.01。这在非科技出版物的地方综合年鉴上,还是可以接受的。比如前述的党员分项合计实例中,实数是“党政机关工作人员 12331 名,企事业单位、民办非企业单位专业技术人员和管理人员 25868 名,工勤技能人员 2963 名,农牧渔民 61319 名,学生 716 名,离退休人员 25980 名,其他职业人员 7066 名”,合计是 136243 名。按前述处理后的数字展开相加,是 136245 名,相差仅为 2。

四是许多精确到个位的统计数并不都是实际意义上的绝对数,有的是抽样调查的结果;有的在计算方法或渠道来源上存在一定的偏差;有的是因测量工具或方法改进后,得出更精确的结果。所以有不少统计数据历年都会做出调整。而这种以万、亿为单位的处理数,看似不够精确却扩大容错空间。比如前述的梧州市行政区域总面积,实际上梧州市总行政区划一直没变,2015 年之前沿用的数据一直是 12588 平方公里,2015 年调整为 12572.48 平方公里。如果历年年鉴一直记录的是 1.26 万平方公里,从某种意义上说这个数据反而更为准确,更为科学。

五是随着大数据时代的来临,从不同渠道获取的数据会越来越多,人们的思维在大数据的冲击下会发生改变,“绝对的精准不再是追求的主要目标,会适当忽略微观层面上的精确度,转向从宏观层面获得更好的认知和洞察力”①。所以,新时代也在要求年鉴及年鉴从业者与时俱进,在更宏观的层面上容纳更大的信息量,以满足读者需要。

2. 比较容易出错的书写

数字书写比较容易出错的是,“9 万(亿)~16 万(亿)”不能写成“9 ~ 16 万(亿)”;“15% ~30%”不能写成“15 ~30%”;“阿拉伯数字‘0’的汉字书写形式有‘零’和‘〇’两种,数字用作计量时,其中‘0’的汉字书写形式为‘零’,用作编号时,‘0’的汉字书写形式为‘〇’;……一般数值不能同时采用阿拉伯数字和汉字数字,如 4000 可以写作‘四千’,不能写作‘4 千’。”②但要注意计量单位不应改动。

(二)关于计量单位的规范

计量单位的采用应当统一按照《国际单位制及其应用》(GB3100 - 93)执行。其中面积的法定单位是以平方米(m^2)为代表的小到平方毫米(mm^2),大到平方千米(km^2)的一系列单位。自 1992 年 1 月 1 日起,国务院批准在统计工作和对外签约中一律使用规定的土地面积计量单位:平方千米(km^2)、公顷(hm^2)和平方米(m^2)。时至今日,我们编年鉴时,还在为用公顷还是用亩纠结不已。究其原因,是“亩”有着庞大的群众基础,人们习惯用“亩产”多少计量农作物的产能,至今都没有合适的量词代替并被大众认可、使用。公顷相对于亩太大,平方米又太小,所以在农业类目中,亩还是随处可见的。《梧州年鉴》现今的做法是,15 亩以上的土地面积一概换算成公顷;在农业类目中用“亩产”记述产能的段落,保持用“亩”作计量单位;“平方米”一般用于建筑面积的计量。

四、数据的完整性是年鉴综合质量的保证

地方综合年鉴要达到比较高的质量标准,还应当注重数据表达是否完整。数据表达的完整性包括两个层次,首先数据内容表达要求全面,其次要求数据是相互关联的。

(一)数据内容全面

从数据的内涵来说,数据资料要具备的完整性,包括数据所反映的对象是否全面和数据本身是否反映了全面的事物两个方面。“数据所反映的对象,特别是有关全局的数据、宏观的数据、整体的数据收录是否全面,如记述工业要有固定资产原值、产品销售收入、利税以及企业留利的数据资料,记述商业要有主要消费零售量、零售额数据等等。”③数据本身是否反映全面,指的是录入数据的统计口径及统计范围是全局的而非局部的。就《梧州年鉴》而言,收录的全局数据应反映全市性的而非市本级的,或者不能只是某个部门的数据。

① 孟小峰、李勇、祝建华:《社会计算:大数据时代的机遇与挑战》,《计算机研究与发展》2013 年第 12 期。
② 《出版物上数字用法》(GB/T 15835—2011),2011 年 11 月 1 日。
③ 王丽君:《重视年鉴中数据资料的运用》,《年鉴信息与研究》1996 年第 1 期。

由于年鉴是逐年出版的连续性刊物,数据的完整性还体现在同一数据及数据组在不同卷次中的连续记载,统计口径应做到统一,注意和上年数据对比,形成时间和空间上的连续性和系列性,“以反映事物的持续变化及结果,保证年鉴中该数据资料的完整、齐全”①。

从数据的表达方式上看,数据要表达全面,还应在年鉴中多方运用列表、曲线图、柱状图等生动、直观、形象的演绎形式,使冗长繁琐的文字记述化繁为简,一扫沉闷的版面布局,增强年鉴的可读性。

(二)数据相互联系

数据是信息的反映,孤立的数据或相互缺乏关联的数据往往缺乏对事物的说明意义,成为使用价值不大的信息。数据的关联性越来越为人们所重视,尤其是进入大数据时代后,更注重相关关系,全面关联的数据通过大数据技术处理成为人们了解世界的更好视角。“年鉴的意义在于对数据进行筛选,经过有序化、系列化的加工处理,使之成为能够说明某一事物的有价值的信息。年鉴中数据的取舍应以其在反映事物中是否为不可缺少的要素为标准,反过来说,数据只有作为说明某事物不可缺少的要素存在时才是有意义的。”②在年鉴中,我们需要的是相关联的全面数据,仅提供某一事物侧面的片断数据是缺乏记载价值的。

五、数据处理技术的提高是保证年鉴升级的重要条件

年鉴内容涉及百科,年鉴编纂工作者即使不能做到博学多才,也应通过努力学习拓展知识面,改善自身知识结构,这是提高数据处理技术水平的基础。因此,年鉴部门有必要对年鉴编纂人员及供稿人员进行针对数据处理的一系列培训。首先,要让他们熟悉与数据处理相关的各种规定和标准,学习和掌握《地方综合年鉴编纂出版规定》《出版物上数字用法》及《国际单位制及其应用》等规定,学习所在省(自治区)出台的相关规定和标准以及所编年鉴自定的行文规范等。其次,请来统计部门或统计专业的行家、教授,传授统计学基本知识和原理,让编纂人员掌握一些常用的统计概念、统计口径、统计公式,熟悉和掌握国民经济和社会发展的各种统计数据。最后,年鉴涉及百科知识,编纂人员还需要学习其他学科的知识,特别是经济学方面的常识。这也应作为培训年鉴工作者的一项重要内容。这样才能掌控各行各业涉及不同领域的数据,培养对数据的敏感度,纠错补缺,做到科学准确的记述。

地方综合年鉴录用的数据必须准确、规范和完整。由于数据来源多样,不可能从一开始就达到要求,需要对数据进行技术处理。要确保数据的准确性必须对数据进行多方核实。为保证数据使用的规范化,数字书写应按照国家标准执行。注重数据内容表达全面和数据之间的相互关联,以编纂质量较高及数据完整的年鉴。

责任编辑:范锐超

① 王丽君:《重视年鉴中数据资料的运用》,《年鉴信息与研究》1996 年第 1 期。

② 杨汉成:《统计数据在年鉴中的规范化运用》,《山西统计》2003 年第 11 期。