年鉴大数据出版之探析

赵瑞红　常晶会*

摘　要　随着我国各行各业信息化的深入发展，大数据的运用越来越广泛。国内年鉴信息化的发展现状表明，现有信息化水平已无法满足为现实服务的需求。相比之下，年鉴大数据对更好地发挥年鉴的存史及使用价值具有较大的推动意义。年鉴大数据的定义以及年鉴大数据标准化加工所需的技术与条件决定了年鉴的传统功能出现了突破和拓展。年鉴大数据的呈现形式则满足了读者多方面的需求，同时也提高了年鉴的开发利用程度。因此，年鉴大数据出版呼之欲出，并且有广阔的发展前景。

关键词　年鉴　大数据　出版

年鉴作为一种重要的战略性大数据知识资源，它如实记录每一年度某个地区或某种行业各方面发展变化的重要事实，连续描述社会发展的历史轨迹、主要脉络和发展机制，对人们客观地认知、研究中国的过去、现在和未来，科学制定和实施发展战略具有重要的知识价值。基于这些研究和思考又会产生新的战略，导致新的变化，可以产生对社会发展有启示的资料，所以具有重要的研究价值。作为连续出版的、反映年度资料的工具书，年鉴的基本属性包括年度性、连续性、资料性和工具性，在此基础上，年鉴的作用与意义体现出它的社会性和现实性。

2015 年 8 月，国务院办公厅颁布的《全国地方志事业发展规划纲要（2015—2020 年）》提出我国地方志事业发展的总体目标："到 2020 年，全面完成第二轮修志规划任务，实现省、市、县三级综合年鉴全覆盖，加快信息化和方志馆建设，做好第三轮修志工作准备，加强对社会修志的指导和管理，基本形成地方志编修体系、理论研究和学科建设体系、质量保障体系、资源开发利用体系、工作保障体系'五位一体'的地方志事业发展综合体系，努力开创地方志事业发展新局面。"①2016 年 12 月，中国地方志指导小组办公室印发的《全

* 赵瑞红，女，山东省梁山县人，《中国学术期刊（光盘版）》电子杂志社有限公司副总经理，主要研究方向为数字出版；常晶会，女，河南省林州市人，《中国学术期刊（光盘版）》电子杂志社有限公司年鉴部副主任，主要研究方向为数字出版。

① 国务院办公厅：《全国地方志事业发展规划纲要（2015—2020 年）》，2015 年 8 月 25 日。

国年鉴事业发展规划(2016—2020 年)》指出:“积极探索‘互联网 +’背景下的年鉴编纂。……加快年鉴信息化建设。将年鉴信息化建设纳入全国信息方志与数字方志建设工程,充分利用已有信息基础设施和数据资源,加快年鉴信息化建设步伐。支持民族地区年鉴信息化建设。逐步建立年鉴数据库。实现国家、省、市、县年鉴资源共享,面向社会提供优质服务。”①2017 年 2 月,中国地方志指导小组办公室印发的《全国信息方志与数字方志建设工程实施方案》提出:以“三网一馆两平台”(中国方志网、中国地情网、中国国情网、国家数字方志馆、地方志综合办公平台、地方志新媒体传播平台)为建设内容,“互联网 + 地方志”利用和发展成为总体目标的重要组成部分。② 年鉴信息化已成为年鉴界的热点话题,在地情网建设、数字方志馆规划、网络年鉴设计、志鉴新媒体尝试等方面,都要不断进行讨论、研究、探索和实践。

一、年鉴大数据的现状

信息化是指充分利用信息技术,开发利用信息资源,促进信息交流和知识共享,提高经济增长质量,推动经济社会发展转型的历史进程。③ 自 2016 年中国地方志指导小组办公室提出,将年鉴信息化建设纳入全国信息方志与数字方志建设工程以来,年鉴界在理论层面与实践层面都进行了尝试和探索,现有的年鉴信息化方式多样、差异性较大。

通过对全国各省的地情网建设情况进行统计,总体形势是:全国除云南省和西藏自治区外,均建设有地情网;地情网上年鉴的呈现形式有网页版、数据库、pdf、doc 等形式,其中,江苏省、浙江省、黑龙江省是以年鉴数据库的形式呈现,可进行条目检索、全文关键词检索等;湖南省和甘肃省仅可查看全文 pdf;安徽省、辽宁省、青海省、河北省、宁夏回族自治区 5 个省份虽然建设有地情网,但目前没有将年鉴发布在地情网上。其他 18 个省份是以网页的形式呈现,可根据栏目导航打开相应的内容;在 18 个以网页形成呈现年鉴的省份中,内容不可复制的包含广西壮族自治区、河南省、江西省、上海市、天津市和重庆市等 6 个省(区、市)。其他省份的年鉴内容则可复制。④

2002 年,李国新提出:“中国年鉴要想在未来仍然作为社会信息资源的组成部分而存在,年鉴的生产者要想使年鉴资源真正为社会所利用,就必须走构建大规模的中国年鉴资源数据库的道路。”在他看来,“这种大规模年鉴资源数据库不能是单种年鉴的简单堆积,也不是把众多单种年鉴归拢到一个网址下就行,而是需要对纳入数据库的年鉴按照数据库的结构、形式、功能进行内容的整合,甚至对原有内容做出深度挖掘和开发,实现信息增值”⑤。从全国各省年鉴的信息化建设情况来看,各自单立,形式各异,内容加工程度不

① 中国地方志指导小组:《全国年鉴事业发展规划(2016—2020 年)》,2016 年 12 月 22 日。
② 中国地方志指导小组办公室:《全国信息方志与数字方志建设工程实施方案》,2017 年 2 月 10 日。
③ 中共中央办公厅:《2006—2020 年国家信息化发展战略》,2006 年 5 月 8 日。
④ 此信息均是对各省地情网站中年鉴的信息化情况逐一进行查看而得出。
⑤ 李国新:《中国年鉴的创新之路:集团化、数字化、网络化》,《年鉴信息与研究》2002 年第 1 期。

同,没有对年鉴进行充分的碎片化处理,资源的整合相对简单,且没有对年鉴的内容进行深层次的挖掘,统一起来相对困难。

目前的年鉴信息化水平仅仅可以满足人们查找年鉴文献的需求,没有达到年鉴进行知识服务所需的水平。只有将年鉴文献内容碎片化,使年鉴资源知识化,从而便于对海量知识进行查询、挖掘和发现,才可为读者提供更直接有效的知识服务与研究工具,即通过年鉴大数据出版的方式来实现。

二、年鉴大数据出版的必要性

(一)年鉴大数据出版的概念

大数据的概念在人们不断的探索和研究中,其内涵和外延日渐丰富。首先,“大数据出版”是一个过程。“大数据出版”是出版者真正理解海量文献数据,深入挖掘各种用户和读者研究和学习的需求,通过出版为读者提供知识服务的过程;①“大数据出版”是一种全新的出版观念和手段,出版者将海量出版物通过技术手段转化为可分析的量化数据,并集成数据库实现信息关联,出版者通过挖掘分析实现信息增值,通过出版发行为读者提供知识服务的过程。② “大数据出版”就是将海量的出版物转化为可制表分析的量化形式,并通过建立数据库使信息产生相关关系的过程。③ 其次,海量的数据与出版物是不可缺少的条件,这些数据和出版物是大数据出版的对象。再次,要对这些海量的数据与出版物进行大数据挖掘。“大数据挖掘”是指有组织、有目的的收集数据,通过分析数据使之成为信息,从而在大量数据中寻找潜在规律以形成规则或知识的技术。④ 如美国学者舍恩伯格就提出“大数据出版”是通过技术手段将出版内容转化为计算机可以检索和运算的信息,并将这些信息集成一个大数据库,实现文本的挖掘和分析。⑤ 最后,大数据出版可以更好地为读者提供知识服务。所谓知识服务是指从各种显性和隐性信息资源中,针对人们的需要将知识提炼出来的过程,它是以资源建设为基础的高级阶段的信息服务。⑥

因此,“年鉴大数据出版”可认为是对海量的年鉴数据进行标准化和规范化加工,并对其进行大数据分析,通过出版为读者提供知识服务的过程。

(二)年鉴大数据的作用与意义

年鉴大数据的作用与意义可以从科研服务、政策研究服务、学习与教学服务、情报服务等四个层面来体现。

---

① 王明亮:《关于“大数据出版”的一些体会和猜想》,《中国新闻出版报》2013 年 8 月 29 日第 5 版。

② 许晶晶:《“大数据出版”对图书馆知识服务的机遇与挑战》,《出版发行研究》2015 年第 7 期。

③ 张振宇、周莉:《“大数据出版”的理念、方法及发展路径》,《出版发行研究》2015 年第 1 期。

④ 谭磊:《New Internet:大数据挖掘》,电子工业出版社,2013 年,第 23 页。

⑤ [美]维克托·迈尔-舍恩伯格著,周涛译:《大数据时代》,浙江人民出版社,2013 年,第 104 页。

⑥ 王明亮:《信息服务到知识服务的转变——对标准化信息与知识服务产业化运作模式的探讨》,《中国高校技术市场》2002 年第 3 期。

1. 科研服务

在科研过程中,逐年可比的材料、该领域最新成果相关材料是不可或缺的。年鉴大数据能够提供年鉴资源的深度挖掘和新知识的发行途径,使得用户更加直观、方便地看到事物连续发展的轨迹,洞察不同事物的相互联系和影响,发现事物相互作用与发展变化的规律,进而支持对各行业、各学科宏观、微观发展规律的研究;支持准确实事信息的获取,为学习、研究提供真实的课题资料。

2. 政策研究服务

年鉴具有资政作用,但此作用并非一本或几本年鉴即可做到的。年鉴大数据能够迅速汇集各种年鉴及其相关文献中的相关政策、法规信息,提供政策研究、决策前瞻的资源参考,帮助政策研究者和制定者明确战略发展目标、准确把握政策实施发展、变化的轨迹和趋势,为科学决策提供依据,提高领导艺术和领导水平。

3. 学习与教学服务

年鉴大数据能够培养学生采集、整理、分析事实信息、发展规律、应用规律的能力;为教学活动提供高质量的素材和手段;为广大学生全面提供有关领域课程的最新内容,培养学者自我教育、自我提高的能力,适应学习型社会对高端人才的要求。

4. 情报服务

数据是最能直观反映信息的情报,而年鉴中存在大量此种具有重要价值的情报。年鉴大数据不仅提供系统、全面的市场数据和相关的行业信息,辅助情报人员进行市场竞争分析,还提供各个地方的资源、交通、投资政策、风土人情等信息,帮助企业进行投资和经济决策。例如,一个企业需要了解养老服务现状及发展趋势,那么需要对医疗卫生机构数、医疗卫生机构床位数、医院病床使用率、65 岁及以上人口数、65 岁及以上人口数占总人口的比重、人口死亡率、人口出生率、总人口数等各种数据进行调研,而这些数据,多数存在于年鉴中,如果利用年鉴大数据,则可以轻而易举完成此数据调研。

要实现以上年鉴大数据的作用,还需要经过一个年鉴大数据出版的过程。

三、年鉴大数据出版的可行性

上述年鉴大数据的出版方式需要有一个前提,即年鉴“生产、传播、扩散”全过程的打通。目前,全国已有多家年鉴编纂单位使用年鉴编纂平台,实现了年鉴编纂从大纲设计、组稿、编纂、校稿、审稿至稿件完结的全程信息化办公。年鉴编纂的信息化可以在年鉴撰稿完结的同时,完成对年鉴内容的碎片化处理和对知识内容的分类管理,进而为后面建立资源数据库和年鉴大数据出版奠定基础。

网络年鉴是年鉴大数据出版的形态之一。目前,《苏州年鉴》《成都年鉴》等都已有尝试。网络年鉴不局限于纸质篇幅,可以更好地与其他相关内容进行链接整合,可以对每条年鉴数据内容进行无限的拓展与丰富,进而增加年鉴内容的丰富性、可读性、趣味性,更好地满足读者需求,提高年鉴资源的开发利用程度。

早在十年前,年鉴大数据出版的研究与实践已经开始。2009 年《中国经济与社会发展统计数据库》(现《中国经济社会大数据研究平台》)面世,它收录了 2302 种统计年鉴、年报、统计报告、普查资料等以及国家统计局发布的月度数据与季度数据,同时整合了行业年鉴中的行业数据及综合年鉴中的统计数据等,通过挖掘和分析这些数据中的统计指标以及这些统计指标的衍生指标,配合相应的计算软件,形成一个能够反映中国社会、经济发展规律,满足读者发现问题、理解问题、解决问题的数据出版物。

基于《中国年鉴网络出版总库》规范的、标准化的以及经过分析和挖掘的年鉴数据,这些年鉴数据按照行业进行分类,形成不同的年鉴专辑,并整合到现有的行业知识服务平台中,与已有的该行业的不同类型的数据资源,如期刊、报纸、博硕士论文、会议论文、图书、标准等以及从互联网中挖掘而来的政务公开数据、新闻、行业解读等内容相结合,形成可以精准满足该行业人员各项需求的知识服务体系。

四、年鉴大数据的出版方式

年鉴大数据研究目前集中于大数据的标准化,对知识性的数据来说,需要从以下几个方面对知识数据进行标准化加工。

首先需要对知识元(最小的知识单元)进行分类。目前的知识元主要分为五个层次——事实、理论、规划、规则和方法。事实类知识元包含现象、事件、工程、实物图片、测量数据及统计数据等;理论类知识元包括概念、思想、原理、理念、假设、公式、理论数据等;规划类知识元包括战略、对策、方针、预测、规划、计划、方案等;规则类知识元包括法律、法规、政策、规章、制度、标准、规范、程序等;方法类知识元包括认知、理论、实验、试验、技术、生产、规程、技巧等。

年鉴的碎片化标准包括形式碎片化标准与内容碎片化标准两种。形式碎片化便是要按文献的结构,将篇、章、节、要点、段落、句子、文献责任人、出处等一点一点的做出来。做出来之后还需要做成一个机器能够识别的内容,这里便需要按照内容碎片化的标准将其进行全文 XML(extensible markup language)结构化处理。XML 即可扩展标记语言,用于标记电子文件使其具有结构性的标记语言,可以用来标记数据、定义数据类型,是一种允许用户对自己的标记语言进行定义的源语言。① 这样就可把每篇文章的每个句子,每个词都标记出来,就可以认识、推理。加工则按照 5W1H 的方式进行加工,5W1H 分别是指现象与行为的主体(Who)、时间(When)、地点(Where)、主题及其类别(What)、意义(Why)和方式(How)。

对所有的对象进行统一的规范标识,不同的人、组织、机构都有不同的名称,不同的人有一样的名字,或者一个机构有不同的名字,这些都需要做统一的规范处理,这样才能把人、事件、机构等之间的关系准确建立起来。例如,目前中国知网在数据加工过程中,对文

① 全国科学技术名词审定委员会:《地理学名词》(第 2 版),科学出版社,2006 年,第 211 页。

献的作者、发文机构、基金等进行规范后，保证读者对文献的查全查准。对刊名进行规范，保证引文数据链接的准确性。

开展年鉴大数据的分析则需要制定事实知识本体—语义的标准，有本体本身，怎么命名，有什么属性，有什么逻辑关系，本体间的关联是什么等。本体之间的关联关系即事物之间的相关联系与作用，包含部属、结构、共生、互斥、因果等，比如人物与组织机构的隶属关系，组织机构之间的关系，思想与决策、决策与执行、行为与效果的因果关系，法律法规、政策与实施效果的因果关系、产业链上下游、相关产业之间的结合关系，不同区域经济社会统计指标的相关关系等。

对年鉴内容进行规范和标准的碎片化加工之后，年鉴数据库的检索功能就不再是过去的文献检索，而是可以进行系统的知识检索、数据挖掘和知识挖掘，同时可以接入各式各样的知识服务平台。所谓系统的知识检索是指通过条目及条目的知识关联一次发现一个完整的知识体系，并通过关联关系，呈现知识网络，支持系统的学习和调研。

2005 年，《中国年鉴网络出版总库》问世，截至 2018 年 2 月 17 日，数字出版的年鉴为 4832 种，37071 册，条目数为 32313985 条。具体如下表所示：

表 1　年鉴收录资源情况一览表

文献类型	国际级		中央级		省级		地市级		区县级		合计	
	册数	种数	册数	种数	册数	种数	册数	种数	册数	种数	册数	种数
年鉴	261	27	5139	541	4640	425	7705	516	10848	1475	28593	2984
统计年鉴	24	4	1004	75	1659	122	2581	194	106	18	5374	413
统计类资料	26	9	429	136	620	289	120	80	14	14	1209	528
综合类资料	105	21	979	356	624	409	133	94	54	27	1895	907
合计	416	61	7551	1108	7543	1245	10539	884	11022	1534	37071	4832

《中国年鉴网络出版总库》检索方式包括条目检索、控制检索、整刊检索、跨库检索等多维专业检索模式，大大提高了检索的精度和效率。整刊检索可按照地域、行业、专辑及出版单位等进行导航，整刊内容分别从年份、栏目等进行呈现，可以查看该年鉴从创刊至今的所有卷册，并通过对所有卷册的所有栏目进行整合和梳理，可纵向查看某个栏目历年的发展脉络，从而帮助读者呈现相对系统的知识脉络。条目可以分为 16 种类型：总结报告、领导讲话、远景规划、各种事件、法律法规、统计公报、统计图表、各种政府文件、标准、人物、论文、大事和其他的作品等。通过对条目大数据进行分类和分析，可对年鉴内容进行相应的分析与挖掘，《中国年鉴网络出版总库》通过对条目内容进行分析与挖掘，形成了国情资源库、区情县情库、公共管理库、图片信息库、机构信息库以及行业频道等年鉴增值信息库，实现了年鉴的整刊出版与条目大数据出版并存。

在此需要提及的是“知网节”功能，知网节是指提供单篇文献详细信息和扩展信息的知识网络节点，它构建了具有多角度知识发现功能的知识网络体系，具有支持知识获取、学习、发现和管理的强大功能。通过“知网节”功能，《中国年鉴网络出版总库》将年鉴内

容与 CNKI 的学术期刊、博士论文、优秀硕士论文、报纸等资源深度整合，可以轻松搜集大量同类资源。

有关专家曾提出一个设想，即如果把“大数据出版”的概念再加以拓展，将其每一条数据“出版”到互联网上更大的大数据云层之中，就是使微数据与云层中的微数据发生“强耦合”，那么，大数据的内容将与整个互联网虚拟社会融合到一起，而无处不在。①

从年鉴大数据出版的概念出发，我们能够想到的更深层次的年鉴大数据出版的方式，不仅仅是像上述出版方式一样，将年鉴条目大数据进行挖掘与重组，而是将年鉴 XML 数据内容进行动态重组。如某个事件或机构的历史变迁，某个人物的历程，可以通过动态重组，直观地展示出来；并且此内容可独立存在，也可与其可拓展的网上的视频、图片、图表内容相关联，进而形成一个更为立体的、更具事实性的、表达力更强的、内容更丰富的年鉴大数据集合，通过将这种年鉴大数据集合与相应的行业知识服务平台相接，便可以更好地进行知识服务。

年鉴作为连续出版、反映年度资料的工具书，其内容的权威性、全面性、科学性、系统性已经历了严格的出版审核。但基于传统的出版模式，其内容存在着包罗万象且表现形式单一等问题。传统的数字化并不能解决此问题，需要转变思维，以碎片化动态重组形式进行结构再造，并加上相关材料以充实内容，将年鉴区域性、行业性、年度性的特点充分表现出来，形成适合网络传播且各行业能够充分利用的知识服务产品，以更好地对历史负责，为现实服务，替未来留迹。

责任编辑：冷晓玲　宿万涛

① 王明亮：《关于“大数据出版”的一些体会和猜想》，《中国新闻出版报》2013 年 8 月 29 日第 5 版。