

古籍数字化工作统筹协调机制的构建^{*}

陈得媛

摘 要 我国古籍数字化工作中的理论认识难以统一,多元主体各自为政,标准不一,选题重复,人才缺乏。在对古籍数字化工作中的乱象与隐忧进行分析的基础上,提出古籍数字化工作统筹协调机制的构建策略,如将古籍数字化上升为国家事业,成立古籍数字化业界联盟,基于元数据统一数据格式,解决版权之争以及培养专门复合型人才等,从而有效协调和解决古籍数字化工作中存在的矛盾和问题,促进古籍数字化工作的可持续发展。参考文献9。

关键词 古籍数字化 统筹协调 持续发展

分类号 G255.1

ABSTRACT There is currently no common conception for the digitization of ancient books in China, but various entities are involved in the work, with different standards, repeated selection of subjects and a lack of skilled personnel. Based on the analysis of the confused situation and potential risks in the digitization of ancient books, this paper provides suggestions for the establishment of a coordinating mechanism for the digitization of ancient books, such as raising its status to the national level, setting up a consortium for the digitization of ancient books, adopting metadata as a uniform format, solving the copyright issues and training specialized personnel, in order to effectively deal with the difficulties and problems and promote the sustainable development of the digitization of ancient books. 9 refs.

KEY WORDS Digitization of ancient books. Coordination. Sustainable development.

CLASS NUMBER G255.1

近30年来,我国古籍数字化工作经历了从最初的摸索尝试、零星制作到如今规模开发的发展过程,取得了令人瞩目的成就,一批大规模、基础性的古籍文献被开发为数字化产品,并成功走向市场,如《文渊阁四库全书》电子版、《四部丛刊》电子版、《中国基本古籍光盘库》、《中国历代基本典籍库》等等。制约古籍数字化实现的一些关键技术,如汉字字符集、文字识别、版面还原和全文检索等问题经过持续的研究和试验,已基本解决。然而,在古籍数字化工作繁荣的形势下,我国古籍数字化工作缺乏全局性的统筹协调机构和协调机制的弊端日渐显露出来。从事古籍数字化工作的多元主体从各自利益出发,各自为政,各行其事,自立标准,古籍数字化工作重复选题,产品莠莠不齐,既浪费了有限的人力物力资源,也给读者带来诸多不便。

这些问题得不到解决,就会影响到我国古籍数字化工作的可持续发展。

1 古籍数字化工作中的乱象

1.1 对于古籍数字化的理论认识难以统一

目前关于什么是古籍数字化,理论层面的认识众说纷纭,分歧很大,其中最主要的观点有四种:①所谓古籍数字化就是利用现代信息技术将古代文献转化为电子媒体的形式,通过光盘、网络等介质保存和传播^[1]。②古籍数字化就是利用数字技术将古籍的有关信息转换成数字信息,存储在计算机上,从而达到使用和保护古籍的目的^[2]。③所谓古籍数字化就是采用计

^{*} 本文系华中科技大学国家哲学社会科学创新基地项目“社会信息科学研究”(批准号:SISI-HUST0801)研究成果之一。

算机技术,对古籍文献进行加工、处理,制成古籍文献书目数据库和古籍全文数据库,用以揭示古籍文献中所蕴涵的极其丰富的信息资源,为古籍的深度开发打下良好的基础^[3]。④古籍数字化是指运用计算机技术,对古籍文献进行加工和处理,建立书目数据库、全文数据库和综合检索系统,并通过光盘、网络等途径进行传播^[4]。除了对古籍数字化定义的理论认识存在分歧外,对其它诸如数据规范问题、汉字标准问题、分类问题等古籍数字化的一些关键问题的理论认识也长期存在分歧,难以统一。

1.2 古籍数字化工作主体多元

从世界范围来说,中国古籍数字化是一项世界性工作,中国大陆地区、台湾、香港以及欧美国家的华人根据所拥有的古籍资源以及技术、资金条件,各自为战,进行相关的古籍数字化工作。其中,国内除一些古籍爱好者和研究者根据自身的兴趣和需要自主进行某些专题的古籍数字化以外,较系统、成规模地开展古籍数字化工作的主要有三类机构:①大学教学和科研机构。这类机构的古籍数字化工作目的性强,主要是根据相关机构的教学和科研需要来选定数字化的古籍和相应的数字化格式。如北京大学中文系的全唐诗检索系统、全宋诗分析系统等。②各类图书馆,尤其是大学图书馆。图书馆的古籍数字化工作主要是根据其馆藏资源有选择有目的地进行,特色鲜明,封闭性较强,而且不连续,不系统。如国家图书馆的敦煌文献、历代拓片数字化项目等。③各种商业性机构。商业机构古籍数字化工作的内容和形式主要是由市场需求或潜在的市场需求来决定,哪一类古籍有市场,就进行哪一类古籍的数字化。考虑到市场运作,商业机构的数字化古籍常常会选择一些大型类书、丛书。如迪志公司和书同文公司开发的《四库全书》、《四部丛刊》,国学公司开发的《国学宝典》等等。

1.3 标准不统一,数字化格式多样

由于缺乏统一的古籍数据库国家标准,目前进行古籍数字化工作的各个机构只是根据自

身的技术特长和需要来设计古籍数据库,因而所采用的数字化技术标准不一致,格式多种多样。从目前已进行数字化的古籍来看,数字化格式除常见的txt、doc、html、超星格式外,还有exe、pdf、wdl、pdg、ebk、edb等20多种,有的还采用位图格式、多媒体格式、图片格式等等。古籍的数字化格式几乎涵盖所有现行的数字化工具。数字化古籍检索平台和检索方式也不统一,主要有access、mysql、sqlsever等检索平台。即使使用同样的数据库平台,也会因为开发商的再次技术加工而人为形成数据格式差异,给数据兼容及随后的跨库检索应用带来极大不便^[5]。

1.4 选题内容大量重复

虽然进行古籍数字化工作的单位不少,但关注的焦点过于集中,选题过度重复。如20世纪90年代大陆、香港、台湾两岸三地的研究机构和出版机构合作,陆续开发了大批的古籍光盘出版物,极大地丰富了中文古籍光盘资源,其中比较有影响力的是《中国基本古籍库》、《四库全书》、《国学宝典》、《四部丛刊》、《古今图书集成》等光盘版古籍,但其中的大量内容是重复的。如文渊阁《四库全书》的数字化就存在着上海人民出版社、迪志文化出版有限公司和武汉大学出版社的三个版本。仅仅是《二十五史》,几乎现在称得上古籍数字化项目的产品都全部或部分包含了它们。而且这种局面导致古籍数字化工作整体布局失衡,一些较为热门的古籍不断被数字化,冷僻一点的则无人问津。

1.5 复合人才缺乏

在古籍数字化过程中,最重要的工作是在数字化之前对古籍进行整理,包括版本的选择、内容的校勘等。目前既具备古籍整理知识和技能,又能熟练应用现代信息技术的高级复合型人才较少。这类人才需要大学设立有关专业来进行培养,但是目前各大学还没有专门为培养古籍数字化工作人才而单独设置专业。

此外,古籍数字化作业内版权之争、技术之争、人才之争屡屡出现,这些都是当前古籍数

字化工作乱象的突出反映。

2 古籍数字化工作乱象背后的隐忧

2.1 因理论认识的分歧导致对古籍数字化工作发展方向认识不明确

任何具体的实践工作都是基于一定的理论认识,不同的理论认识决定着实践工作的方向和形态。因此,对什么是古籍数字化的理解不一致,使得相关古籍数字化工作的目标、数字化的内容和提供给社会的产品各不相同。如基于古籍数字化就是利用现代信息技术将古代文献转化为电子媒体的形式,通过光盘、网络等介质保存和传播这种定义的古籍数字化工作,就不可能在数字化古籍的研究支持功能上下功夫,其主要工作就是输入和扫描。从目前读者使用的情况看,这些以键盘输入或扫描形式形成的网络版、光盘版古籍作品,还有一些使用阅读器仅供阅读的古籍资源,只能说是实现了古籍资源阅读的计算机化,并不能称为真正意义上的数字化古籍产品。另外,古籍数字化中诸如数据规范问题、汉字标准问题、分类问题等理论认识长期存在分歧而难以统一,表面上看是一个学术问题,而在实际操作层面,它往往被嵌入到制度环境或市场环节中。古籍数字化的一些重大理论如果长期争论下去,不能形成统一,人们就会对古籍数字化工作的发展方向难以做出正确的选择和判断。

2.2 由古籍数字化工作主体多元化而形成各自为政

近30年来,由于从事古籍数字化工作的主体众多,各主体的利益诉求不尽相同,因而其所进行的古籍数字化工作性质和目标千差万别。

科研机构和图书馆组织的古籍数字化建设,通常是以项目形式申报,在整个项目流程中,会受到来自所属组织及机构内部的古籍资源、业务基础、募资能力、人员素质、管理效率乃至工作风格等多种因素的影响,很多时候,这些因素可以归结到行政组织的某些固有特性上,数字化工作而不得有所掣肘。因此,往往某

些具体技术问题,如工作平台的选取、工程进度的安排等,都出现多方博弈。

对于商业运作的公司而言,制度环境相对简单,技术力量也较学术单位强大,但他们在项目过程中,每一步都会有成本—收益的考虑。以市场为导向,以营利为目标,这是商业公司的性质所决定的。而当学术目标与商业利益发生冲突的时候,学术目标和社会责任往往被放在次要位置,以致这些商业公司推向市场的古籍数字化产品质量参差不齐。更有一些商业机构在进行古籍产品开发时,忽视了古籍数字化工作的基本社会责任,数据库错误严重。有的商业机构只是把古籍文本进行文字扫描导入,疏于版本选择和校勘,古籍电子文本错误百出,难以阅读,造成严重后果。

值得注意的是,这些不同利益诉求的古籍数字化工作主体由于利益驱使,各自独立研发,相互保密,数据库互不兼容,导致资源不能共享。如目前使用的三套最主要的字符集内码互不兼容,基于不同字符集内码开发的数字化成果只能在各自的平台上运行,形成了数字化古籍内容基本相同,而资源却不能共用的局面。

另外,由于没有统一的协调机构,各主体分散作业,至今未能完整了解哪些古籍已经被数字化,哪些古籍正在进行数字化,更无法知晓其数字化格式和利用程度。这种局面,一方面造成大量的重复建设,一方面又使这些学术资源无法充分被利用^[6]。

2.3 格式标准不统一,并形成技术垄断,降低了古籍资源的兼容性和可用性

开展古籍数字化工作的根本目的是实现全社会乃至世界范围的资源共享,但由于各古籍数字化工作主体的规模不同,发展水平不同,能力不同,采用的数据格式、浏览器模式也不同,至今也没有建成一个真正联合统一综合编目的数字系统。目前正在进行或已经完成的古籍数字化产品中,由于缺乏统一标准,对数据资源的描述方式不统一,文献语标规范不统一,数据库结构内容不规范,缺乏规范的控制系統,缺乏统一的检索端口和阅读平台,致使全国已有的数

字化古籍资源各自独立、互不兼容,读者检索起来很困难,这也使得古籍资源被大量重复数字化。

更重要的是,古籍数字化缺乏统一的强制性标准,已在事实上形成了个别数据公司或集团的信息资源垄断。有些公司采取技术措施控制图书馆向其购买信息资源的访问期限,或者信息资源文档缺少通用格式,相应公司提供的数字化文档资源必须用对应的诸如超星、书生或方正 Apabi 格式阅读器才能读出来。这种信息垄断行为,大大降低了古籍资源的兼容性与可用性。

2.4 古籍产品开发缺乏连续性

古籍数字化是多方协调与合作的系统工程,但目前我国古籍数字化基本上是拥有古籍资源的图书馆与拥有数字化技术的公司之间的简单合作,从古籍数字化项目选题开始,就缺少多方互动。从目前已经完成或正在进行的古籍数字化选题方式上看,基本上是依赖开发者自己的意愿,或者依赖文献资源拥有者的倡导,或者依赖部分学者的评议,甚至依赖权威的断言。其实,古籍数字化最根本的一点就是项目开发应该根据“什么最需要”而非“是否有价值”,如果考虑到开发公司自身的生存问题,还应该加上是否能够赢利或具有赢利希望的原则去选定,而专家学者通常着眼于项目“价值”选题,公司着眼于是否赢利选题,恰恰缺少用户对古籍资源“需要”的介入,致使那些被数字化的古籍缺乏连续和深度开发的激励机制^[7],在提供有关古籍内容本身科学、准确的统计与计量信息,提供与古籍内容相关的参考资料、辅助工具的研究支持功能上停滞不前^[8]。

3 古籍数字化工作协调机制的构建策略

为了使古籍数字化这一知识发展服务能够可持续发展,迫切需要合理构建全局性的统筹协调机制,有效协调和解决古籍数字化工作中存在的矛盾和问题。

3.1 将古籍数字化工作上升为国家事业

应尽快成立像国家古籍数字化工作委员会一类的协调机构,主动负起责任,统筹全局,协调各方利益关系,制订中长期的古籍数字化战略规划。我国的古籍文献浩繁,分布在不同系统和部门的图书馆及其他机构。因此,古籍数字化建设需要各部门相互协调,相互配合,统一规划,统一指挥,避免重复浪费。各个机构建立的古籍数据库,虽然为学术研究做了一些贡献,但由于各自目的不同,数据库的设计思路也不同,各数据库之间不能兼容已经妨碍到数据库进一步发展。另外,一个具体机构也不可能建立一个包罗万象的古籍数据库。因此,这些工作只能由国家有关部门在全国范围内进行资源配置,组织人力物力来完成。

3.2 成立古籍数字化业界联盟

由国家古籍数字化委员会牵头成立一个由科研单位、图书馆、专业开发公司及有相关经验的团体、个人组成的古籍数字化业界联盟。在这个业界联盟的组织下,对古籍数字化工作的现状进行普查^[6],制订相关的技术标准和规范,引领行业发展,避免重复建设。例如,古籍数字化建设中现有字符集不够用是一个瓶颈,无论是 GBK 还是方正超大字符集,其中臆造出来的毫无用处的字符太多。这种情况主要是由于设计方不知道实际需求。GB2312 的 6000 多个汉字可以满足一般文献的需要,但对于处理古代典籍到底需要哪些字符,没有古籍数字化的经验,很难有清晰的认识。业界联盟的建立,可以集中经验,共同制定一个相对稳定、合理的字符集标准。在文献信息处理过程中,每一步都将涉及标准问题,如异体字的统一、信息的组织分类、叙词表(关键词)的确定等环节。

3.3 基于元数据统一数据格式

这里所说的统一,并不是把所有古籍文献内容统一为文本或图像格式,在现有条件下,把全部古籍电子格式文本化也不现实。由国家古籍数字化工作委员会牵头,依靠业界联盟,基于

目前图书馆界和国外学术资料信息化的普遍经验,将资源对象的语义信息统一为元数据格式。这种基于元数据的统一数字化格式通过建立国家数据中心,从而制定古籍数字化建设的统一标准就可以实现^[5]。

3.4 协调各方面的利益关系,解决版权之争

当前困扰古籍数字化建设的一个重要难题就是版权问题,说到底就是利益关系问题。传世古籍除了今人的影印、点校、注释、翻译之外,是没有著作权的,不涉及版权问题。但数字化以后,这个问题就变得复杂了。数字化算不算古籍整理?有没有产生新的著作人、版权人?又是否能够保证使用者所引用资料的正确无误?图像版的古籍有无版权?如果有,版权是归藏书单位,还是归图像制作者?国家尚未出台一部针对数字化古籍中这些问题的法律法规,导致古籍数字化中有关机构常常在技术上设置壁垒,以保证自身的权益不受侵害,而这些壁垒恰恰与古籍数字化实现资源共享的终极目标相矛盾。因此,必须构建运行有效的协调机制,从全局的高度公正妥善协调各方利益关系,使各古籍数字化机构在维护自身法定权益的前提下,允许制作有关的古籍数据库,以利于学术研究。如国家有关部门可以主动与相关机构协调,亦可抽出部分分散投入到各课题中的资金来补偿有关机构,实现双赢^[9]。

3.5 设立古典文献与计算机技术学科和专业,培养古籍数字化方面的专门复合人才

人才培养是个复杂的系统工程,需要教育主管部门、学校和社会共同努力。就目前来说,数据库开发主要由计算机专业的人才来进行,专业的局限对古籍数据库的建设有一定限制。要统筹解决这一问题,应该考虑在有条件的学校中打破原有招生专业,设立一个文理交叉学科——古典文献与计算机技术学科和专业,这个专业以培养古典文献素养较好、计算机技术出色的从事古籍数字化工作的高级专门人才为

目的,或者在已有的图书馆学专业和古典文献专业有意识地吸收计算机专业人才攻读硕士学位。

此外,针对古籍数字化工作,在学术评价机制、国家项目资金投入机制、财政税收机制等方面适当变革,协调配合,共同推动我国古籍数字化工作健康持续发展。

综上所述,我们必须正视当下古籍数字化工作中的乱象,尽快构建国家级的统筹协调机构和协调机制,统筹全局,协调处理各方利益关系,通过政府、各机构组织及企业、个人及国内、国际间的通力合作,保证古籍数字化工作的可持续发展。

参考文献:

- [1] 陈阳. 中文古籍数字化的成果与存在问题[J]. 出版科学, 2003(4): 47-48.
- [2] 彭江岸. 论古籍的数字化[J]. 河南图书馆学刊 2000(2): 63-65.
- [3] 王桂平. 我国古籍数字化的现状与展望[J]. 图书情报知识, 2000(4): 50-53.
- [4] 吴家驹. 中文古籍数字化的进展与主要成果述评[J]. 南京师范大学学报, 2004(3): 178-183.
- [5] 唐磊. 整合古籍数字化资源的必要与可能[N]. 中国社会科学院院报, 2007-09-18.
- [6] 李明杰, 肖秋惠. 中国古籍数字化资源调查与分析[J]. 图书馆杂志, 2002(5): 25-28.
- [7] 徐青, 石向实, 王唯. 古籍数字化资源的深度开发[J]. 图书情报工作, 2007(3): 95-97.
- [8] 李国新. 中国古籍资源数字化的进展与任务[J]. 大学图书馆学报, 2002(1): 21-26.
- [9] 彭国忠. 古籍数字化后带来的新问题[N]. 中国文化报, 2008-03-02.

陈得媛 华中科技大学图书馆助理研究员。通讯地址: 湖北省武汉市珞瑜路 1037 号华中科技大学图书馆。邮编 430074。

(收稿日期: 2009-03-18)