DOI:10.13530/j.cnki.jlis.2023008

数字人文应用服务中的数据版权风险及防范策略

欧阳剑

摘　要　数字人文研究中的人文数据及其衍生数据产品具有较高的学术价值,随着“数据资产化”理念的出现,数字人文应用服务中的数据版权风险逐渐突显出来。数据版权风险是数字人文应用服务体系构建面临的现实挑战之一,也是影响数字人文健康发展的关键因素。本文从数字人文应用服务中的数据建设版权风险与数据使用行为风险入手,对各种来源数据的权属、授权许可方式、用户使用行为等潜在风险进行了分析,从法律法规、数据合法性、开放许可协议、应用服务平台等角度对数字人文应用服务中的数据版权风险成因进行了分析,并针对风险来源及风险成因提出了防范策略。现有法律法规对数据版权的界定比较模糊,数字人文应用服务面临的一些新的问题需要通过完善法律法规来解决;需要确保数字人文应用服务中人文数据的合法性,从源头上杜绝版权风险;依据合理使用原则和例外条款,主动规避可能产生的数据版权风险;通过增强数字人文应用服务平台的安全性和审核机制降低侵权风险。参考文献 29。

关键词　数字人文　应用服务　人文数据　数据版权

分类号　G251

Risks and Prevention Strategies of Data Copyright in Digital Humanities Application Service

OUYANG Jian

ABSTRACT

In recent years, digital humanities research has gradually received widespread attention, and data development and sharing in digital humanities application services have provided various support for humanities research. The construction of humanities data in digital humanities research has become a hot spot. With the emergence of the concept of “data assetization”, the value of humanities data and its derivative data products in digital humanities research has been highlighted. There are many copyright infringement risks in the development and utilization of data in digital humanities, and the copyright risks of data in digital humanities application services are gradually emerging. Data copyright risk is one of the real challenges faced by the construction of digital humanities application service system, and it is also a key factor affecting the healthy development of digital humanities.

This paper starts from the data construction copyright risk and data infringement behavior in digital humanities application services. Firstly, it analyzes the ownership, authorization mode and copyright risk in the construction process of various source data such as commercial data, self-built featured data, research institution data, user-generated content based on the network, and crowdsourcing data in digital humanities application services. Secondly, it also analyzes the potential risk sources such as user usage

通信作者:欧阳剑,Email:oyjjj@163.com,ORCID:0000-0001-5867-2852(Correspondence should be addressed to OUYANG Jian,Email:oyjjj@163.com,ORCID:0000-0001-5867-2852)

behavior in digital humanities application services, and analyzes the causes of data copyright risk in digital humanities application services from the aspects of laws and regulations, data legality, open license agreement, and security platform of application services. Corresponding prevention strategies are proposed for the sources of copyright risk and the causes of data copyright risk. The existing laws and regulations are vague in defining data copyright, and some new problems faced by digital humanities application services need to be solved through the improvement of laws and regulations. It is necessary to perfect the laws and regulations on data copyright. At the same time, in the process of constructing humanities data, it is necessary to ensure the legality of humanities data, eliminate copyright risks from the source, and actively avoid potential data copyright risks in accordance with fair use principles and exceptions. In terms of the use of humanities data, the infringement risk can be reduced by enhancing the security and audit mechanism of the digital humanities application service platform. 29 refs.

KEY WORDS

Digital humanities. Application services. Humanities data. Data copyright.

0 引言

近年来,数据驱动的人文学科研究日益受到关注,人文数据及其衍生数据产品显现出了巨大的价值,数字人文研究中的人文数据建设成为热点。数字人文应用服务平台作为数字人文基础设施应用服务的重要组成部分,更加注重数据开放服务实践,以为人文研究提供更多的支撑。与此同时,人文数据建设与数据应用服务中所涉及的数据版权问题也逐渐突显出来。数字人文中的人文数据来源渠道多元化,权属关系复杂,在数据的开发、利用、服务过程中存在一定的版权侵权风险。由于没有统一、明确的数据版权界定,数据提供方的权益难以被合理维护。而在数字人文应用服务过程中往往容易忽视对数据版权的保护,在数据使用与传播的过程中容易发生侵权行为,给权益人造成损害。

数据版权风险是数字人文应用服务体系构建面临的现实挑战之一,也是影响数字人文健康发展的关键因素。本文在分析数字人文应用服务中的版权风险来源和成因的基础上,有针对性地提出防范策略,为促进人文数据在更大范围内的共建共享提供参考。

1 从数字版权到数据版权

1.1 从数字化服务到数据化服务

数字人文应用服务平台是一种支持人文领域科研活动的基础设施,是数字环境下开展人文研究的基本条件,包括与研究主题相关的文献、数据、软件工具、学术交流和出版的公用设施及相关服务等[1],也是开展数字人文服务的主要方式。数字人文应用服务平台与传统数字图书馆服务平台的显著区别在于它是以数据化的数据为主,并辅以数字化研究工具,为人文学者提供一个数字化的研究环境。总的来说,数字人文应用服务是以“数据化”为主要方式对各种类型的人文资料进行组织与揭示,将分散于展览馆、图书馆、档案馆、博物馆等文化机构的数字化馆藏资源进行整合,作为服务的基础数据,通过平台化的架构为人文学者提供时空、文本、社会关系等各类量化分析与可视化数据服务。随着越来越多的人文数据相互融合,基于数据组织的多维度数据平台应时而生,逐渐由传统图像扫描型资源库向数据化数据库转变,以满足人文学者日益增长的研究数据需求。

1.2 从数字版权到数据版权

数字版权是指作者及其他权利人对其文学、艺术、科学作品在数字化保存、复制、传播方面依法享有的一系列专有性的精神权利和经济权利的总称,诸如复制权、数字发行权、修改权等。它的作用主要体现在作品的确权、维权、授权和版权交易等方面,既包括作者利用新兴的数字版权管理技术与版权法规保护原创智力成果免受侵权,也包括传播者对其传播的作品依法享有的专有权利。随着数字化环境的发展,数据驱动型研究使人们越来越清楚地认识到数据的价值,尤其在数字经济时代,数据版权的重要性逐渐显现。数据版权是数字版权的一种延伸,数字网络环境以数据获取、存储、管理和分析为主要对象,数据价值衍生出新的内涵和形式。数据价值显著增长,其中蕴含了引人瞩目的经济利益,增加了数据价值的"需保护性"[2]。相对于现有法律来说,数据版权的规定还有待规范与完善。

1.3 数据版权的挑战

数据是数字人文研究的基础和核心之一。数字人文研究中的人文数据由计算机处理的可计算化的数字形式编码,主要由格式化数据、文本、图像、音频和视频等组成[3]。人文数据跟科学数据及互联网大数据相比也存在明显的不同,除了新兴的原生人文数据,更多的人文数据来自于传统文献(一般不包含创作过程)的数字化,或是对原始数据进行开发或加工而产生的分析数据或汇编数据,与著作权有着天然的联系。在数字人文研究中出现了数据库、数据内容、数据软件等新的版权内容的数据化形式,人文数据经过加工、再利用会衍生出新的价值,其中蕴含了一定的经济利益,推动人文数据从"数据化"逐步向"数据资产化"迈进。由于人文数据的处置方式有复制、传播、利用等,数据资产化使得新的数据版权问题凸现。在此背景下,数据成为数字人文应用服务中版权保护的主要对象,数字人文应用服务平台也面临新的数据版权风险,正如王兆鹏教授于2018年11月在"大数据与人文地理信息数据库建设"国际会议上所提醒的,"涉及对上传数据者拥有的知识产权的法律保护也是需要考虑在前面的议题"[4]。2022年3月26日在"古籍智能信息处理"系列研讨会第二讲中,专家们也就古籍数据的相关版权问题进行了热烈的讨论[5]。人文数据的知识产权问题逐步引起了数字人文研究者的重视与关注:为合理地建设人文数据并开展数字人文应用中的开放数据服务,如何使遵守数据版权保护要求与合法使用、有效传播人文数据达成二元平衡,成为亟待解决的关键问题。

近年来,随着人文学者对数字人文基础设施建设的重视,数字人文应用服务平台的研究成果不断增多,新的应用服务平台也不断出现,如上海图书馆历史人文大数据平台[6]、山东大学的东亚数字人文平台[7]、中国古籍基础数据应用服务平台[8]等。数字人文应用服务平台中的数据版权问题也开始引起学者关注。数字人文应用服务平台以数据服务为核心,其数据来源复杂,不同的数据有各自的许可协议与授权方式,对于使用与管理来说极易造成侵权。而且随着数字人文应用服务中文献资源"数据化"的开展,对文献资源的二次组织与重构工作也面临着一些版权界定方面的新困难,极易出现版权纠纷。而应用服务过程中也存在许多版权风险,如果没有统一、明确的数据版权界定,数据提供方的权益难以被合理维护。因此,数据版权问题也在一定程度上影响着数字人文应用平台开展数据服务,有必要从用户数据需求与版权人利益的角度出发,探究数字人文应用平台数据开放服务中的版权风险问题。

2 数据版权风险来源

数字人文应用服务平台是以实现数据开放与重构为目的的数字人文基础设施。不同于数字图书馆支撑平台,数字人文应用服务平台更注重数据共享和使用,利用多元化的数字技术

实现数据分析与知识发现,向人文学者提供数据与知识服务。数字人文应用服务中的版权风险主要来自人文数据建设和应用平台服务两个方面。

2.1 人文数据建设中的版权风险

数字人文应用服务中的人文数据主要来自文献原始数据或汇编数据。传统的数字资源是人文数据的重要来源,通过对传统的数字资源组织与重构并进行文本化、数据化处理,实现数字资源向人文数据转变。人文数据建设过程中的版权风险主要存在于两个方面。

2.1.1 多源数据带来的版权风险

数据化过程中的数字资源主要来源于外购商业数据库、自建数据库、研究机构的数据以及网络开放获取数据等。由于这些数据的权属、授权及许可方式不同,在对其进行统一处理时或可造成侵权风险。

(1)商业数据库具有专业性强、权威性高的特点,其中一些商业数据库成为数字人文应用服务中的主要数据来源。近年来随着数字人文的推广,越来越多的商业公司(如中文在线、知网等)推出了一种新的数据服务形式——数字人文应用服务。由于认识上的不足,一些图书馆与学术机构在订购商业数据库时容易忽视数据库商对于数据的授权与许可。目前,大多数数据库采购方案中只提及访问权或使用权,但未进一步说明该权利包含的具体内容[9],在引进合同中也没有明确的关于数据使用范围、使用方式等条款。同时,由于数据库存在不同的使用权利(包括访问权、存档权、使用权、复制权、改编权等),图书馆和学术机构通常取得的是对数据库的访问权与使用权,甚至于有的数据库商本身对其授权的数据不享有版权或不享有授权给第三方的权利,因而,被授权方(图书馆和学术机构)在使用数据库时可能存在侵权风险。

(2)自建特色数据库是数字人文应用服务中的重要数据来源之一。多年来,随着数字化项目的开展,文化机构已拥有规模庞大、种类繁多、具有高价值的数字资源。数字化的文档资料、数据库和检索系统等数字学术资源逐渐成为人文研究的基础平台[1]。自建特色数据来源本身比较复杂,一般以自有版权的特色资料为主,但自有版权存在多种版权形式[10],还有部分自建数据是对不同馆藏类型的原始纸本资源(如馆藏图书、报纸、期刊论文等)进行加工的结果,如果对原始资源的使用超出了《中华人民共和国著作权法》(以下简称《著作权法》)规定的少量、适当的范围,则可能产生侵权风险。近年来,越来越多的特色数据是通过对商业数据库中的数据进行二次加工组织与重构、提炼而成,这涉及版权所有者的复制权、传播权、财产权等一系列权利,处理不当可能会产生侵权风险。

(3)研究机构是一类特殊的团体,其研究一般得到相关基金资助,因而其研究数据作为研究成果的一部分,版权一般受到资助机构的约定,如美国国立卫生研究院、英国惠康基金会、英国研究理事会、德国科学基金会等机构请求或要求课题组把受其资助出版的文章存放在公开数据库内[11],通过在线下载、离线共享或定制服务等方式向社会开放共享,在使用数据时需要注明数据来源,或通过许可的形式进行重用,否则就会产生侵权风险。目前,研究机构的部分数据在开放存取(Open Access,OA)期刊上发布,作为作品的副产品,这类数据大多采用开放使用的知识共享署名许可协议(CC BY 协议)授权[12]。根据 CC BY 协议,作者拥有版权,所有用户只需给予作者及作品来源适当的认可,即可下载、摘录、再用、存档和分发作品[13]。CC BY 协议可确保数据被广泛利用,一般不涉及著作权问题,版权风险相对较少,而对于少部分数据还需要跟研究机构签署书面授权,否则可能面临侵权风险。作为文史类研究机构的代表成果,中国历代人物传记数据库(CBDB)、中国历史地理信息系统(CHGIS)等也有明确的版权使用限制[14],即其数据仅限于使用在非商业的学术研究和教育上,而且非授权不得以任何电子载体形式或通过因特网下载的方式重新发布

数据。

(4)基于网络的用户原创内容(UGC)极大丰富了数字人文应用服务的数据建设,成为一些数字人文研究的数据来源,如维基百科。用户参与创作的内容大多采用GNU自由文档许可协议(GNU Free Documentation License)及知识共享许可协议(CC协议),特别是CC协议在推动知识共享与鼓励大众创造知识的同时充分考虑了作者的权益,可以在使用条款中注明对用户原创内容的允许行为与限制行为,从而选择性地保留了部分作品版权。这种版权保护方式在促进用户创造知识的同时也保证了作品完整性与作者署名权,用户原创内容除作者本人申明之外,非商业性使用一般不涉及著作权问题,但如果作者保留部分作品版权,那么在使用过程中也可能会出现侵权行为。

(5)政府开放数据为数字人文提供了丰富的人文数据。近年来,随着各级政府开放数据政策的实施,多个领域的不同规模的数据集,特别是文化领域的数据集,得以向公众开放。目前政府开放数据的使用以免费为主,北京[15]、上海[16]、深圳[17]等地的政府数据开放平台分别在其法律声明、使用条款、服务条款等规定中对用户的权利与义务进行了约定,即用户有免费使用、传播分享和利用及再利用数据资源的权利,但不得有偿转让平台上的各种数据资源。作为以科研为主要任务的数字人文应用服务,使用政府开放数据不会存在侵权风险,但在侵犯任何第三方的合法权益的情况下则会有风险。

(6)数字人文研究中的众包数据是对人文数据的有益补充。在新的数字网络环境下,以人文数据建设为目的的数据众包活动不断增多。数据众包活动可根据数字人文项目的需要,采用大众共建的方式,实现定制化的数据获取以及数据加工方案设计与执行服务,为数字人文项目提供标准化、结构化的可用数据[18],从而推进了人文学科资源的数据化。在出版模式上,由众包活动产生的资源,有的以发起者的名义出版,而有的则是以集体创作出版。由于数字人文中的数据众包实践还处于新兴事物阶段,部分数据众包发起者的成果版权意识停留在传统的个人/集体著作模式上,导致项目成果权利归属不明晰,或漠视众包参与者的权益。目前还没有完善的众包数据共享政策与法规,因此,数字人文应用服务在使用众包数据时会存在一定的数据侵权风险。

2.1.2 人文数据建设过程中的版权风险

数字人文研究中的人文数据建设除了收集、整理原生数据之外,还有人文数据复原与人文数据重构两种形式[3],即对传统文献资源进行转录、改编、重组。转录属于人文数据复原操作,即经过识别、清洗、转换等一系列加工处理工作构建人文数据集,把原始文献的数据按原有结构转换成数据格式,并按照原始文献的知识体系重建原来的系统化数据与知识结构。改编、重组属于人文数据重构,即按照人文学者的研究需要,以研究课题的数据结构进行组织与重建,并根据知识之间的关系建立起人文数据之间的关联;改编、重组主要是摘取原始文献中的主要事实和数据,或选取原始文献中的篇章、事实或数据等并对其进行有机排列,其中文献选编、年鉴名录等原始文献是这一过程的重要处理对象。

对传统文献资源进行转录、改编、重组等工作与《著作权法》中的使用权、复制权、改编权等密切相关。当符合《著作权法》所允许的少量、适当引用他人作品的规定,且在自主加工成数据的过程中投入智力劳动,从而具有独创性,一般不涉及著作权问题,版权风险相对较小。然而《著作权法》对原始的、非结构化的数据的版权保护比较薄弱,因为一些原始数据可能不符合"原创性作品"的创造性要求,而通过对原始数据进行开发或加工后产生的分析数据或汇编数据才被视为著作权客体,现有关于数据的法律保护也存在空白。但在有些情况下,法院愿意将版权保护范围扩大到涉及数据且具有足够创造性的作品[19],因此,在人文数据建设的过程中很容易因为版权不明确而产生纠纷。

2.2 数字人文应用平台服务过程中存在的版权风险

数字人文应用服务为人文学者提供了更便捷的数据访问方式,但同时也带来了数据侵权的风险。目前数据服务普遍缺乏版权风险监测与评估机制。数字人文应用服务平台很容易忽视数据库商对于数据的授权许可协议,在签署引进合同时没有明确限定数据的使用范围及方式,容易忽略履行注意义务,在用户对数据库中原始内容进行复制、修改、汇编等操作时没有明确的侵权提示,造成数据使用方式、使用范围不当。此外,评估版权风险也是数字人文应用服务的一项重要工作,如果在数字人文应用服务的过程中,应用平台未对用户使用服务数据的行为进行合法性审核,则也有可能承担共同侵权风险。

尽管数字人文应用服务在数据保护方面采取了一定措施,但仍存在用户利用系统漏洞恶意非法批量下载数据或借助第三方软件非法访问数据的侵权行为。数字人文应用服务平台为用户上传数据提供了空间,但因开放的特殊性,平台作为数据的开发与共享网站,对用户上传侵权数据的行为,具有未承担起审查用户上传数据是否侵犯他人著作权的义务的过失,属于放任侵权数据存在、逃避其应负的与其所享有的权利与利益相匹配的义务,根据《著作权法》第四十七条和第四十九条之规定,平台应当承担相应的侵权责任。

3 数据版权风险成因分析

3.1 数据版权法律法规不完善

随着新的数字环境的出现,数据驱动型研究使人们越来越清楚地认识到数据的价值。但数据版权是数字经济兴起后产生的新事物,数据版权难以在现有保护对象中找到自己的明确定位,实现法律赋权仍任重道远。现有法律法规对数据版权的界定比较模糊,以前的著作权法难以适用于新的数据版权,比如,对商业数据修改、汇编、重组等操作不符合现有法律对“原创性作品”的创造性要求,难以将版权保护范围扩大到涉及数据且具有足够创造性的作品。新环境下,对原始数据进行加工处理后产生的衍生数据、用户交互过程中所产生的数据等的版权归属难以界定,数据众包实践产生的众包数据的版权归属模糊,这些都使得数据版权存在风险。另外,支持从海量数据中发掘有用信息的文本和数据挖掘(Text and Data Mining,TDM)技术已被越来越多地应用于文化机构,TDM 的必要步骤是对大量作品或数据的复制,但我国目前已有的限制与例外制度无法为文化机构的 TDM 行为提供合法性支撑[20],使数字人文应用服务面临很大的版权风险。

3.2 数据合法性不明

数据建设是提供数字人文应用服务的前提。数据来源的多样化导致了数据本身合法性不明、版权归属复杂等问题。自建数据一般版权归本机构所有,在进行数据建设时可以直接获得本机构的授权,而且关于本机构的成员成果的版权问题也比较容易解决,能够较好地避免版权纠纷。但商业数据库具有明确的版权要求,而且不同数据库有不同的使用权利。在人文数据建设与服务过程中,对商业数据进行加工、汇编等处理时很容易忽视数据库商对于内容的授权与许可,或超出许可使用范围而出现侵权行为。研究机构数据及开放数据均有比较多的版权限制,而且在使用的过程中因授权协议的不同也极易产生侵权风险。此外,数据建设过程中对传统资源的二次组织与重构行为本身的合法性也难以明确。

3.3 开放许可协议多样、标准不一

对研究机构及网络上的开放数据的使用大多需要得到 GNU 自由文档许可协议、CC 协议等开放许可协议的授权,同一许可协议因授权要素的不同而有不同版本,不同授权版本的许可协议对数据许可利用的规则也有所差别。来源

多样的数据集在使用许可协议上难以统一,一方面增加了数字人文应用服务中的多源数据组织、建设过程中处理不同协议兼容的难度;另一方面,多源数据中的不同授权协议的适用范围、适用地区、权利许可方式、权利限制均有所不同,使得协议误用的风险增加,从而也增加了侵权风险。

3.4 平台应用服务中的风险

数字人文应用服务平台是数字人文应用服务中的重要设施,是人文数据与用户连接的桥梁,数据使用的合法性通过应用服务平台得以体现,同时,平台的安全性也是影响数据版权风险的一个重要因素。首先,上文提到的数据来源的合法性、版权归属等是数字人文应用服务平台建设的主要影响因素,而目前人文数据建设在数据来源的合法性与版权使用方面存在风险。其次,数据的使用方式、使用范围、数据的安全性等也是数字人文应用服务的重要影响因素,数据版权保护与人文数据开放之间存在一定的矛盾,而利益平衡是版权保护的根基。对于研究者来说,他们希望能够方便、快捷地获取自己需要的人文数据,可以对内外部数据进行整合与关联访问,并得到"一站式"的数据服务,但这种高度开放的数据服务具有较高的侵权风险;而对于数字人文应用服务来说,应将数据版权保护作为平台建设与提供服务的主要考虑因素。再次,数字人文应用服务过程中也存在不少的数据侵权行为,如何通过技术手段保护数据版权依然面临挑战。

4 数据版权风险防范策略

4.1 健全数据版权法律法规

数据驱动背景下人们逐渐意识到数据的经济价值与法律属性,目前专门针对数据的法律法规基本上以对"人"的权利保护或保障数据安全为基础,如欧盟《通用数据保护条例》(*General Data Protection Regulation*,GDPR)、我国 2021 年 6 月通过的《中华人民共和国数据安全法》等。我国现有法律对数据版权保护不足,比如我国刑法第二百一十七条是对著作权的保护,但数据版权在其保护对象中的对应对象并不确切[2],同时,鉴于 TDM 的重要价值和意义,我国有必要在法律层面对此类行为进行明确定性,以建立成熟完善的数据法律规范,这是保护数据版权的基础。

数据流通的整个环节会涉及各种版权问题,尽管对于数字版权基础理论的讨论已经很多,但实际的数据版权分析案例相对较少,不足以指导信息政策与法律法规的制定。通过对以往数字版权的具体案例分析可以总结出"思想观念""原创性""表达方式"等重要概念,由于数字人文的数据二次组织与重构模式涉及诸多复杂情况,这些基本版权概念是制定数据版权相关法律法规的重要依据。对人文学者基于数据组织与重构所形成的智力成果的版权保护性质与汇编作品相类似,在传统著作权法强调思想与表达二分法的原则下[21],也就是只保护数字内容的原创表达方式。无论是用户对于已购数字内容的复制与转载,还是对数据的变更与删改,这些都打破了传统著作权模式,相关法律法规有必要根据数字环境的变化对版权保护的对象做出相应调整,对一些条款加以修订与补充,有必要制定基于处理版权数据集所产生的衍生数据的法律规范,对变革型使用的规定进行重新思考,明确合法访问和利用数据的方式和范围,为数据的开发与共享提供法律保障。

4.2 确保人文数据的合法性

数字人文研究中的数据来源渠道多样,因此,在应用服务平台上开放提供数据服务之前,应确认数据本身的合法性、版权归属等,对数据是否合法、数据内容的权利归属是否存在问题进行审查,从数据来源上保证数据权属清晰,是数字人文应用服务平台的基本义务。对于具有许可约定的要遵守其协议约定,对于有版权使用限制的要设置特定使用用户、特定的

访问范围,对于商业数据库中存在版权风险的内容要进行采购及许可授权前的预审,避免出现版权上的风险。目前,版权数据所有者一般采用数据购买或许可授权的方式参与数字人文应用服务平台的建设,如中国台湾"中央研究院"在其"中华文明之时空基础架构系统"中所使用的《中国历史地图集》,被版权所有者中国社会科学院独家授权其制成数字化(矢量化)地图[22]。

对自建数据则要厘清数据来源的合法性,特别是要审查来自对商业数据库选择、编排、再组织而成的数据是否符合《著作权法》第十五条对汇编作品的规定,并且在对具有版权的文献资料的数据化过程中不能侵犯原作品作者的著作权,纸质文献数据化过程中也不得侵犯原始权利人的其他权利。研究机构的数据虽然在政策上可以共享,但如果对数据进行二次加工、转换则可能违反其共享权、改编权、非商业性使用权等规定,比较好的办法就是获得相关研究机构的授权,比如,上海图书馆就中国历代人物传记资料库的使用获得了哈佛大学燕京学社的授权。开放数据的授权则比较复杂,如 OA 数据可能含有开放数据共用协议、知识共享许可协议等,它们对开放共享许可的权限也不相同,以 CC BY 协议为例,在搭配非商业用途(NC)、禁止演绎(ND)和相同方式共享(SA)三个授权元素下,可分为 CC BY-SA、CC BY-NC-SA、CC BY-NC 等六种形式。因此,需要清楚许可协议的授权范围,以便做出合法的开放数据使用行为。

完善授权许可机制也是确保人文数据合法性的一种方法。数据许可方与被许可方可以就数据使用过程中的使用方式、使用范围、使用限制、义务等问题依法自愿达成协议。该协议应明确关于许可数据的使用范围、数量、使用目的、具体使用方式等授权使用条款和使用限制条款,以及被许可方享有的权利性质。数字人文中的应用服务需要遵守双方的许可协议,在许可规定的范围内使用数据。

4.3 明确合理使用原则和例外条款

合理使用是在遵守著作权法规定的情况下,以正当性的使用目的为前提,在合理保护范围内无需经过著作权人的许可和同意,也无需给付著作权人任何报酬和费用就可以使用先前作品的行为[23]。

合理使用主要是平衡著作权专有利益与社会公共利益,各国著作权法对合理使用都有规定,如我国《著作权法》第二十四条对合理使用进行了明确的规定;美国在《版权法》第一百零七条中以是否为商业目的或者非营利教育目的、被使用部分与整个作品的比例是否适当、利用结果对著作潜在市场与现在价值的影响作为是否合理使用的条件;我国《信息网络传播权保护条例》(2013 年修订)第七条规定图书馆等可以不经著作权人许可,通过信息网络向本馆馆舍内服务对象提供本馆收藏的合法出版的数字作品,但不得直接或者间接获得经济利益,也就是说只要符合合理使用原则就不会有版权风险。《著作权法》第十五条也规定汇编若干作品、作品的片段或者不构成作品的数据或者其他材料,对其内容的选择或者编排体现独创性的作品,在不侵犯原作品的著作权的情况下,其著作权由汇编人享有,也就是说在人文数据建设中可以将两个以上的作品、作品片段中的某些数据进行选择、汇集、编排成新的数据,这种汇编数据行为符合《著作权法》的规定。

随着数字环境的发展,内容分析研究逐渐兴起,数据挖掘授权方式也开始出现,为平衡权利人与使用人之间的利益,欧盟给数据挖掘的版权保护开设了例外[24]。Elsevier 于 2014 年更新了数据挖掘政策[25],允许研究人员在开展非商业性研究时通过商业数据库提供的应用程序接口(API)对授权资源进行数据挖掘,产出结果可以在"署名—非商业性使用"(CC BY-NC)的 CC 许可模式下进行利用。Springer[26]、SAGE[27]、BMJ[28]等都发布了文本与数据挖掘政策,对文本与数据挖掘进行了规范,在开展非商业性研究时允许研究人员通过其

所在机构进行文本和数据挖掘,在遵守商业数据库协议的基础上非商业性的合理使用数据不存在版权侵权风险。Hathitrust 充分利用文本内容挖掘策略,采用“数据胶囊”(Data Capsule)的方式有效地解决了图书文本内容的版权保护与公众服务的平衡[29]。作为研究性的数字人文应用服务,主要以教育和科学研究等非商业使用为主,数据收集、加工、保存和传播具有非营利性,数字人文中的数据服务应充分利用这些合理使用原则与例外条款,在合理使用制度下规避版权风险。

4.4 确保数字人文应用服务平台的安全性

作为数字人文应用的重要方式,服务平台承担着数据版权保护的法律任务,除了需要审核数据本身的合法性外,数据的使用方式、使用范围及数据的安全性等也需要通过应用平台来实现。因此,可从技术上对数字人文应用平台的数据进行一定的保护,从控制数据库的访问、控制数据的后续使用以及保护数据的完整性与安全性等不同的途径进行版权管理。

服务平台中的技术安全直接影响数据的安全性和可用性,用户的恶意复制、下载行为对版权所有者的利益造成威胁。不断发展的数据保护技术,可为数据版权保护提供了一系列方法和手段,如授权验证与下载、付费订阅等。数字技术同样可应用于数字人文应用服务平台的建设中,通过提出解决方案来进一步提高这些系统的可靠性和安全性,如采用编码、数字水印等方式来防止数据被侵权。数据加密技术在限制非法获取版权数据的同时也给人文学者的使用带来一定的困难,因此,在数字人文应用服务平台的建设中有必要推动人文研究中版权数据的合理应用,以便有效地平衡各利益方之间的复杂关系,实现人文学者所需的数据开放获取环境。

数字人文应用服务需要在应用服务与版权保护中寻找到平衡点,在开放环境为主导模式的基础上增设一部分高质量的数据付费版块,通过建立合理的数据交易机制以强化人文学者的数据版权意识,从而为数字人文应用平台的建设提供更合理的数据开放策略。应用服务平台可充分借鉴 Hathitrust 的“数据胶囊”模式构建数据环境[29],即将受版权保护的内容加以封装,采用特定的算法对封装的原始数据进行分析,通过文本挖掘等方式获取对原始文本的分析结果,禁止用户直接访问内容,避免用户直接接触到原始数据。这种运作机制既可以打破数字版权的屏障,满足人文学者的研究需求,又较好地保护了原始数据的版权。

作为用户共享平台,数字人文应用服务平台的部分人文数据由用户自主上传,在这部分数据面临版权纠纷时,平台应当采取一定的规避风险的措施。我国《信息网络传播权保护条例》制定了网络版权领域的“避风港”规则,即在发生数据版权纠纷时,平台可采取“避风港”原则,通过履行“通知—删除”义务的方式避免承担法律责任。同时,平台应监管用户上传内容,采取删除、屏蔽、断开链接等必要的技术措施来约束用户的不良行为,对其权利加以限制,以保护学者对人文数据开发与共享的利益。

5 结语

数据版权风险是数字人文应用服务体系构建面临的现实挑战之一,也是影响数字人文健康发展的关键因素。数据驱动下的数字人文应用服务以数据建设为重心,数据来源的多样化使得数据的权属、授权及许可方式不同,成为数字人文应用服务中的主要版权风险来源。现有法律法规对数据版权的界定比较模糊,数字人文应用服务面临的一些新的问题需要通过完善法律法规来解决;同时,需要确保数字人文应用服务中的人文数据的合法性,从源头上杜绝版权风险,并依据合理使用原则和例外条款,主动规避可能产生的数据版权风险,确保数字人文应用服务平台的安全性。

致谢:本文系国家社会科学基金项目“数智时代中国古籍基本知识表示及构建研究”(编号:22BTQ070)和中央高校基本科研业务费项目“数字人文视域下的多语种数据平台构建及应用方法研究”(编号:2021114010)的研究成果。

参考文献

[1] 刘炜,谢蓉,张磊,等. 面向人文研究的国家数据基础设施建设[J]. 中国图书馆学报,2016,42(5):29-39.(LIU W,XIE R,ZHANG L,et al. Towards a national data infrastructure for digital humanities[J]. Journal of Library Science in China,2016,42(5):29-39.)

[2] 莫洪宪,胡骞. 技术变革背景下从数据价值保护迈向数据版权保护——以侵犯著作权罪的构成要件补足为进路[J]. 出版广角,2018(15):6-9.(MO H X,HU Q. Moving from data value protection to data copyright protection in the context of technological change:taking the complementary composition elements of copyright infringement as a way in[J]. View on Publishing,2018(15):6-9.)

[3] 欧阳剑,彭松林,李臻. 数字人文背景下图书馆人文数据组织与重构[J]. 图书情报工作,2019,63(11):15-24.(OUYANG J,PENG S L,LI Z. Organization and reconstruction of library's humanities data under the background of digital humanities[J]. Library and Information Service,2019,63(11):15-24.)

[4] 徐笛薇. 中国古典文献数据可视化研究兴起,学者:不能是“花架子”[EB/OL]. (2018-11-24)[2020-06-23]. https://www.thepaper.cn/newsDetail_forward_2665626.(XU D W. The rise of data visualization of Chinese classical literature. Scholars:not a “flower shelf”[EB/OL]. (2018-11-24)[2020-06-23]. https://www.thepaper.cn/newsDetail_forward_2665626.)

[5] “古籍智能信息处理”系列研讨会第二讲[EB/OL]. (2022-03-30)[2022-05-05]. http://www.ai.pku.edu.cn/info/1086/2117.htm.(The second seminar in the “intelligent information processing of ancient books” series[EB/OL]. (2022-03-30)[2022-05-05]. http://www.ai.pku.edu.cn/info/1086/2117.htm.)

[6] 上海图书馆历史人文大数据平台[EB/OL]. [2023-02-01]. https://dhc.library.sh.cn/.(Digital humanities platform of Shanghai Library[EB/OL]. [2023-02-01]. https://dhc.library.sh.cn/.)

[7] 山东大学的东亚数字人文平台[EB/OL]. (2020-08-12)[2020-12-19]. https://xinwen.wh.sdu.edu.cn/info/1003/27924.htm.(East Asia digital humanities platform at Shandong University[EB/OL]. (2020-08-12)[2020-12-19]. https://xinwen.wh.sdu.edu.cn/info/1003/27924.htm.)

[8] 中国古籍基础数据应用平台[EB/OL]. [2020-12-19]. http://121.201.35.124:88/.(Chinese ancient books basic data application platform[EB/OL]. [2020-12-19]. http://121.201.35.124:88/.)

[9] 孙锐,杨新涯,魏群义,等. 文献资产元数据仓储建设关键问题研究——以重庆大学图书馆为例[J]. 大学图书馆学报,2018,36(2):18-24.(SUN R,YANG X Y,WEI Q Y,et al. Study on the key issues of data warehouse construction:taking Chongqing University Library as an example[J]. Journal of Academic Libraries,2018,36(2):18-24.)

[10] 象牙塔中遇到的数字版权问题[EB/OL]. [2020-08-19]. http://www.ncac.gov.cn/chinacopyright/contents/4509/334205.html.(Digital copyright problems encountered in the ivory tower[EB/OL]. [2020-08-19]. http://www.ncac.gov.cn/chinacopyright/contents/4509/334205.html.)

[11] LARIVIÈRE V,SUGIMOTO C R. Do authors comply when funders enforce open access to research?[EB/OL]. (2018-10-24)[2020-12-19]. https://www.nature.com/articles/d41586-018-07101-w.

[12] Creative Commons. Attribution 4.0 International(CC BY 4.0)[EB/OL]. [2020-07-01]. https://creativecommons.org/licenses/by/4.0/.

[13] PLOS 编辑与出版规定[EB/OL]. (2014-10)[2020-06-26]. https://journals.plos.org/plospathogens/s/editorial-and-publishing-policies-chinese#loc-studies-sponsored-by-specific-funders.(PLOS editorial and publication regulations[EB/OL]. (2014-10)[2020-06-26]. https://journals.plos.org/plospathogens/s/editorial-and-publishing-policies-chinese#loc-studies-sponsored-by-specific-funders.)

[14] CHGIS 版权声明[EB/OL]. [2020-12-19]. http://yugong.fudan.edu.cn/views/chgis_copyright.php.

(CHGIS copyright notice[EB/OL]. [2020-12-19]. http://yugong.fudan.edu.cn/views/chgis_copyright.php.)

[15] 法律声明[EB/OL]. [2020-08-26]. https://data.beijing.gov.cn/gywm/mzsm/index.htm. (Legal statement[EB/OL]. [2020-08-26]. https://data.beijing.gov.cn/gywm/mzsm/index.htm.)

[16] 使用条款[EB/OL]. [2020-08-26]. https://data.sh.gov.cn/view/footer-nav/index.html?nav=use-terms. (Terms of use[EB/OL]. [2020-08-26]. https://data.sh.gov.cn/view/footer-nav/index.html?nav=use-terms.)

[17] 深圳市政府数据开放平台服务条款[EB/OL]. [2020-08-26]. https://opendata.sz.gov.cn/maintenance/forward/toTermOfService. (Shenzhen government data open platform terms of service[EB/OL]. [2020-08-26]. https://opendata.sz.gov.cn/maintenance/forward/toTermOfService.)

[18] 岑炅莲,欧阳剑,曾辉. 数字人文项目中的数据众包运作策略研究[J]. 图书与情报,2020(5):125-132. (CEN J L,OUYANG J,ZENG H. Research on humanities data crowdsourcing strategies in digital humanities[J]. Library & Information,2020(5):125-132.)

[19] U.S. copyright protections for market data[EB/OL]. (2018-02-12)[2020-10-20]. https://www.natlawreview.com/article/us-copyright-protections-market-data.

[20] 袁锋,徐琢. 新技术环境下图书馆限制与例外条款的问题与完善研究——兼论《信息网络传播权保护条例》的修订[J]. 图书馆杂志,2022,41(5):31-38,55. (YUAN F,XU Z. Research on the problems and improvement of library limitation and exception clauses in the new technology environment:on the revision of *Regulations on the Protection of Information Network Communication Right*[J]. Library Journal,2022,41(5):31-38,55.)

[21] 王凤娟,刘振. 著作权法中思想与表达二分法之合并原则及其适用[J]. 知识产权,2017(1):87-92. (WANG F J,LIU Z. The merger principle of the dichotomy of ideas and expressions in copyright law and its application[J]. Intellectual Property,2017(1):87-92.)

[22] 《中华文明之时空基础架构系统》版权声明[EB/OL]. (2002-09)[2022-05-05]. http://ccts.sinica.edu.tw/copyright.html. (Copyright of Chinese civilization's space-time infrastructure system[EB/OL]. (2002-09)[2022-05-05]. http://ccts.sinica.edu.tw/copyright.html.)

[23] 吴汉东. 著作权合理使用制度研究[M]. 北京:中国人民大学出版,2013:1. (WU H D. Study on copyright fair use system[M]. Beijing:China Renmin University Press,2013:1.)

[24] 宋昕哲. 欧盟给数据挖掘的版权保护开设例外[N]. 中国科学报,2019-08-06(08). (SONG X Z. EU makes exceptions to copyright protection for data mining[N]. China Science News,2019-08-06(08).)

[25] SHILLUM C. Elsevier updates text-mining policy to improve access for researchers[EB/OL]. (2014-01-31)[2020-08-01]. https://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers.

[26] Text and data mining at Springer Nature[EB/OL]. [2020-07-26]. https://www.springernature.com/gp/researchers/text-and-data-mining.

[27] Text and data mining on SAGE Journals[EB/OL]. [2020-07-15]. https://journals.sagepub.com/page/policies/text-and-data-mining.

[28] BMJ text and data mining (TDM) policy and licence[EB/OL]. [2020-06-27]. https://www.bmj.com/company/legal-information/terms-conditions/legal-information/tdm-licencepolicy/.

[29] JAIMIE M,JACOB J,TIM C,et al. Towards publishing secure capsule-based analysis[C]//Proceeding of the 17th ACM/IEEE Joint Conference on Digital Libraries. Toronto,Canada,2017:261-264.

欧阳剑 上海外国语大学图书馆、上海外国语大学数字学术中心研究馆员。上海 201620。

(收稿日期:2022-05-10;修回日期:2022-11-10)