

# 网络文献的内涵界定与管理

韩娜<sup>1,2</sup> 马海群<sup>1</sup> 张春春<sup>1,3</sup>

(1. 黑龙江大学信息管理学院, 哈尔滨 150080;

2. 黑龙江科技大学计算机与信息工程学院, 哈尔滨 150027;

3. 黑龙江省图书馆, 哈尔滨 150090)

**摘要:** [目的/意义] 针对网络文献快速发展、内涵不断延伸的问题, 现有的网络文献定义存在局限性, 故对网络文献进行重新界定。[方法/过程] 通过对文献和现有网络文献定义的梳理, 结合大数据时代下网络文献的特点和内涵, 抽象出网络文献的定义并将其作为文献进行研究。[结果/结论] 对网络文献给出新的定义, 并根据该定义对现有网络文献进行分类汇总, 进而得到网络文献诸多特征。针对网络文献的特殊性, 分析了网络文献管理中保存和利用现状, 并就知识产权保护、文献存档、检索效率和信息挖掘问题提出解决思路。作为文献和信息技术发展结合的产物, 网络文献的研究有助于扩大文献学研究范畴, 为文献信息学提供新的研究视角。

**关键词:** 网络文献 网络文献定义 分类 文献管理

**分类号:** G250

**DOI:** 10.31193/SSAP.J.ISSN.2096-6695.2021.02.03

随着科学技术的快速发展, 网络应用逐渐融入社会各个领域。“网络文献”从1984年“情报网络文献检索讲习班”中第一次出现, 到现在随处可见, 其内涵发生了巨大的变化。而“网络文献”是否属于“文献”以及是否归属于“文献学”研究的范畴, 迄今为止并没有明确的界定。本文将从文献定义出发, 厘清网络文献是否满足文献要素, 给出网络文献的界定和内涵, 进而分析网络文献在当前整理利用中存在的问题及解决思路。

## 1 网络文献的发展及定义

### 1.1 文献的定义

“文献”这一概念在不同时期其涵义不同, 从而产生诸多不同定义和见解。“文献”一词,

[作者简介] 韩娜 (ORCID: 0000-0001-7530-6459), 女, 讲师, 博士研究生, 研究方向为文献挖掘、文本处理, Email: 17627877@qq.com; 马海群 (ORCID: 0000-0002-2091-7620), 男, 教授, 博士生导师, 博士, 研究方向为信息政策与法律, Email: mahaiqun@sina.com.cn; 张春春 (ORCID: 0000-0001-9371-3891), 女, 副研究馆员, 博士研究生, 研究方向为公共文化、信息服务, Email: 12418713@qq.com。



最早出现于《论语·八佾》，传统解释是：“文”为“文章”，“献”为“贤才”或“贤”。“贤才”的释义随历史流逝而逐渐淡化，大多古代学者均将“文献”认同为“典籍”。现代对文献的定义集中在文献应具有的价值和内容、形式等组成要素方面。如文献是“具有历史价值和认识作用的、以文字记录形式存在的资料”<sup>[1-2]</sup>；文献是“具有使用价值、历史价值的”具体材料<sup>[3]</sup>；周启付<sup>[4]</sup>认为“文献是用文字、图形、符号、声频、视频等技术手段记录人类知识的一种载体，或称为固化在一定物质载体上的知识”；洪湛侯<sup>[5]</sup>提出“凡是用文字表述的具有历史价值和科学价值的图书资料，就是‘文献’。”上述定义将文献是否具有价值及价值类型作为衡量文献的标准。

而中国国家标准《文献著录总则》(GB3792·1-83)将文献定义为“记录有知识的一切载体”。傅广荣<sup>[6]</sup>建议将该定义修改为“文献：记录有知识的一切物体”。朱建亮<sup>[7]</sup>辨析了文献定义的内在和外在表现后，将其定义为：文献是以字符、声像等为信号的，以便于长期保存和广泛传播的物体为信道或载体的人类精神信息的固态品。林申清<sup>[8]</sup>将文献划分为三个要素，分别是：知识（情报）、记录手段或方法、载体。于鸣镛<sup>[9]</sup>以列举方式将文献定义为“兼具人为性、意识性、可读性、传播性、存贮性、符号性、隐值性、物质性和双重性的固态化精神产品”。陈界<sup>[10]</sup>将文献归结为四大元素，并定义文献是“将知识信息用文字、符号、图像、视频、音频等记录在特定载体上的整合体”。可见随着技术发展，文献的功能和社会作用出现新的特质，其概念也有了新的发展。以上定义将文献内容和载体作为要素进行说明，而文献内容也不再局限于知识，而是扩展为“精神信息”和“知识信息”，即知识不仅仅局限于客观知识，也应包含主观知识或主观经验。

朱娜娜<sup>[11]</sup>将文献载体、形式、功用等加以悬置，提出现象学视角的文献定义，得出：文献是以一定形式记录和传递信息的介质，这里的“信息”是“信息链”的统称，包括了事实、数据、信息、知识和智能（情报）。该定义将文献内容从“知识”扩大为“信息”，强调文献与形式和内容无关，更关注文献的内涵，符合当前文献类型不断增加、文献内容不断丰富、文献表现形式多样化的特点。

从朱娜娜提出的文献定义出发，“网络文献”可以视为：采用现代计算机技术手段将文字、符号、图像、声频、视频等多种形式的信息记录在硬盘上，并通过网络进行信息的传递以方便用户使用的介质，符合文献的定义范畴，可以将其视作文献的一种。

## 1.2 网络文献的发展历程

在文献发展过程中信息载体随社会发展和技术水平不断提高而变化。经历了原始社会的甲骨、金石，商代的简策，战国的帛书，直到汉代的纸张，文献的记录和保存才变得普及，从而促进文献发挥其社会功能。19世纪40年代开始，出现了以胶片、胶卷保存的微缩文献和存储在磁、光、电介质上的电子文献<sup>[12]</sup>，也称为数字文献。到20世纪70年代，网络技术蓬勃发展，才出现了“网络文献”这一说法。同时期，类似的叫法还有“虚拟文献”“虚拟资源”“网络出版物”等不同概念。但究其根本，“虚拟”是相对于之前载体的现实表现的具象化而提出的，但并未真正明确“虚拟”的具体含义。故“网络文献”的叫法更能明确传输、存贮、利用的载体是网络，而得到业界的认可。



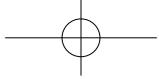
文献载体有三个重要的变革阶段, 分别是纸质文献、电子文献和网络文献。纸质文献是造纸术、印刷术后存在时间最长的文献类型, 未来很长一段时间内也将仍然是文献的主要类型; 电子文献主要包括微缩文献、音像文献和电子出版物, 在发展初期主要是纸质文献的数字化表示而已, 故也被称为数字文献; 网络文献的定义和类型则没有明确的界定。由于电子文献和网络文献的发展时间和表现形式有重合, 在很长一段时间内, 造成二者的概念混淆和互相替换使用情况时有发生。

“网络文献”是科学技术发展的必然产物, 其表现内容也会随新的科技产品出现而变化。王东波<sup>[13]</sup>认为, 网络文献是在计算机互联网上运行的、用户可检索利用的、有历史价值或参考价值的图书资料。但随着多媒体技术的发展, 文献不仅包含图书资料, 还包含音频、视频等形式, 故该定义缩小了文献的范畴。相丽玲<sup>[14]</sup>认为网络文献是在 Internet 上存贮、检索、利用或传递信息的过程中, 通过各种网络传输协议和计算机程序组织起来的信息单元或信息集合。黄晓斌<sup>[15]</sup>将网络文献定义为在因特网上以数字化形式存贮的各种文献信息内容的集合, 并列举了如网络书目数据库、电子期刊报纸、学术会议论文等类型。刘佳<sup>[16]</sup>将网络文献概念确定为: 是依附于计算机存储设备中, 可在网络上传输、识别、利用的, 可稳定获取, 在一定时间内可固定访问的信息单元。以上三种定义都认为网络文献要由网络进行信息内容存贮, 但并没有区分信息内容的第一载体是纸质文献还是电子文献或是网络。郭伟等<sup>[17]</sup>辨析了电子文献和网络文献在载体、可变性、著录规则、检索方式等方面的差异, 提出网络文献是指以计算机网络为第一载体, 通过网络存储, 传递图、文、声、像等信息, 并在网络终端进行联机检索, 以网页的形式再现的信息资源或信息集合。该定义明确网络文献第一载体应是网络, 但信息资源或集合的呈现形式不一定是“再现”, 也可能是直接呈现, 如现在的各种直播, 以网络为第一载体产生、传输、存贮, 可以在终端进行检索, 应符合网络文献的定义范畴。刘石等在关注大数据时代下传统文献学转型过程中, 认为“数字文献是以数字代码形态存在, 依赖计算机系统存取和传输的文本、图像、音频、视频等文献”<sup>[18]</sup>, 但其主要指代的是“重要文献的扫描、保存和开源共享”, 即仍将数字文献视为纸质文献的数字化。综上可见, 网络文献一直没有形成符合当下实际情况的定义。

### 1.3 网络文献的定义

综合已有的网络文献定义或概念, 结合当前科技发展产生的各种网络新鲜事物, 笔者认为, 网络文献是网络原生且通过网络存贮、传递和呈现的信息或其集合。

所谓“原生”指网络文献的第一载体为网络, 信息资源可以在网络上首发, 或与纸质版本同时发布, 只是发布途径不同而已。网络原生是将网络文献与纸质文献的数字化做有效区分的重要指标。网络文献的信息资源类型与文献类型相匹配, 即包含图像、文字、音频、视频等大类及其混合体, 如视频文献中除了声音和图像, 可能还会包含部分文字内容, 很难将现在的多媒体文献单独归结为某一独立类型, 更多的是多种类型的结合使用。网络文献在网络上首发, 可以通过网络进行传递, 其保存方式更多地采用云存储的方式, 便于文献的访问和利用。这个定义弱化了保存、传递、呈现的具体实现方式, 究其原因是科学技术仍在迅猛发展。过去的单点硬盘数据库保存已进化为分布式云端存储, 而计算机网络技术也从过去的有线网络访问发展到今天的移动互联网访问, 5G 技术、人工智能、物联网等新兴技术的普及将带来新的技术变革风暴, 所以, 实现



技术并不是规范网络文献范畴的充分条件,而是其必要条件。至少从目前计算机发展前景来看,不会根本性地推翻网络文献的产生、传递和存贮是基于网络的这一特质。

## 2 网络文献的范畴及特点

《世界互联网发展报告 2020》<sup>[19]</sup> 蓝皮书提到,全球互联网用户数约为 45.4 亿,普及率达 59%,而美国和中国的互联网发展居 48 个受调查国家的前两位。We Are Social 与 Hootsuite 合作发布的《2020 全球数字报告》显示:数字、移动和社交媒体已成为全世界人们日常生活中不可或缺的一部分<sup>[20]</sup>。互联网规模和技术的发展让网络文献的形式不断更新。从本文提出的网络文献的定义出发,可以对目前符合定义的信息及信息集合进行分类,将其纳入网络文献的范畴。

### 2.1 网络文献的范畴

网络文献是网络原生的,即网络文献的第一载体必须是网络。将传统文献的类型分类与网络文献呈现方式结合,研究者将首发于网络或网络与其他载体同时刊发的文献整合划分为以下 9 类:

#### (1) 网络期刊

网络期刊又称为电子出版物、网络出版物。任何在网络上首发的以电子形式存在的期刊广义上都可以称为网络期刊。目前我国网络期刊来源主要有三类:一是期刊社开发有自主知识产权的网络期刊,期刊的采编、发行、传播等都在网络上完成;二是依靠主办单位网站或其他合作网络实现网络化的期刊<sup>[21]</sup>;三是通过权威信息代理机构如中国知网、万方等代理机构进行期刊网络发行。与传统期刊相比,网络期刊可以融入图像、文字、声音、视频等动态内容,还可以为用户提供超链接等互动元素,其延展性强,随着移动互联网的发展,支持在各种个人终端上进行阅读。

#### (2) 网络图书

广义上理解,以网络为媒介存在的各种类型读物都可以视为网络图书。而狭义上来看可以将其限定为电子图书,且根据网络文献定义将其规范为第一发表媒介为网络各类书籍。主要指网络文学,即由作者在网络上首次发表的文学形式。

#### (3) 网络报纸

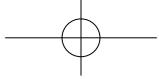
网络报纸指的是在互联网上建立一个独立的网站,其内容可能包含全部或部分纸质报刊内容,但所占比例不高,更多的是提供其他的信息服务,如中国的《人民日报》、美国的《华盛顿邮报》均在网站上提供传统纸质报纸的扩展内容。

#### (4) 电子公告服务

电子公告服务指互联网上的电子信息系统,包含电子公告牌、电子论坛、网络聊天室、留言板等交互形式<sup>[22]</sup>。从内容上可以分为综合类和专业类,综合类电子公告内容侧重于各类政治时事等国内外重大事件新闻信息;而专业类电子公告主要由政府、企业或个人运营维护,发布运营者提供的相关信息。

#### (5) 计算机程序

计算机程序是由专门的程序设计语言编写、运行于某种目标计算体系结构上的代码化指令序



列或符号化语句序列。尽管计算机程序的运行结果是提供某种功能, 即目标结果是一连串的二进制“0, 1”序列, 但其源程序可以视作静态文本, 即文献。越来越多的计算机程序率先在网上发布, 用户可以通过免费或付费的方式进行获取、利用。从整个过程来看满足网络文献的基本定义, 可以视为网络文献的一种。

#### (6) 预印本

预印本是指科研工作者的研究成果还未在正式出版物上发表, 出于和同行交流目的, 自愿通过互联网发布的科研论文、科技报告等。越来越多的机构提供预印本服务, 如美国的 arXiv<sup>[23]</sup> 电子预印本文献库包含物理、数学、计算机科学等领域的学术性信息; 由全球不同国家的志愿者共同建立的 RePEc 网站, 主要提供经济学相关研究成果的论文、期刊文章以及应用软件; 中国预印本服务系统是以提供预印本文献资源服务为目标的实时学术交流系统, 包含国内预印本和国外预印本服务两个子系统; 奇迹文库是为中国研究者提供免费、稳定的预印本平台, 并提倡开放获取<sup>[24]</sup> 的理念, 其主要提供科研文章、综述、学位论文、讲义及专著的预印本; 中国科学论文在线系统<sup>[25]</sup> 采取文责自负的原则, 收录自然科学、教育学、管理学和经济学方面的学术性论文。

#### (7) 社交媒体

社交媒体 (Social Media) 指互联网上基于用户关系的内容生产与交换平台, 是人们彼此间用来分享意见、见解、经验和观点的工具和平台<sup>[26]</sup>, 现阶段主要包括社交网站、微博、微信 (公众号)、博客 (学者个人网站、QQ 空间)、论坛 (在线社区、游记、攻略等)、播客 (在线直播), 以及具有社交属性的各类提供特定类型文献的平台, 如网易云音乐、爱奇艺等等。社交网络的信息 (SNS) 因为其时效性, 被视为记忆工程的组成部分, 如美国国会图书馆建立社交媒体档案库, 用于保存包括 SNS 在内的多种社交媒体记录<sup>[27]</sup>。

#### (8) 网络帮助文件

网络帮助文件是指针对网络资源的学习及应用, 在网上提供的各种资源应用说明、使用手册或应用工具类文档。由于其用途的特殊性, 可以单独划分为一类。

#### (9) 网络数据库

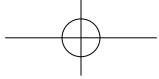
数据库是“按照数据结构组织、存储和管理数据的仓库”<sup>[28]</sup>, 本质上是存储在计算机中的数据集合。因常规数据库并非一定是网络原生, 故在归类时仅将网络数据库归并到网络文献中。网络数据库是指跨越计算机设备在网上创建、运行的数据库, 例如: 分布式数据库、云数据库、物联网数据库、P2P 数据库、移动数据库等。

## 2.2 网络文献的特点

网络文献的来源广泛, 类型多样, 形式复杂, 但均具有以下共同特点:

(1) 网络文献内容丰富, 信息量大。互联网规模的急剧扩大, 使得网络文献信息量急剧增加, 网络文献几乎覆盖政治、经济、科技、文化、教育、娱乐等所有领域, 同时也促进了全球文献信息资源的共享和交流, 而信息的共享又进一步刺激网络文献信息的蓬勃发展。

(2) 网络文献更新速度快, 时效性强, 生命周期缩短。网络文献的产生、交流和获取都在网络上进行, 而网络的传输速度决定其较传统文献缩短了信息时滞, 无空间和时间的限制, 几秒的时间就可以完成网络文献信息从产生到获取的全过程, 极大地缩短了网络文献的生命周期, 加快



了网络文献的新陈代谢速度，网络文献的稳定性降低。

(3) 网络文献来源和形式多样化，规范性相对较差，质量良莠不齐。网络文献来源复杂，既有政府、专业机构发布的信息，又有微博、微信、网络小说等以个人为单位产生的网络信息，既有传统出版机构网络正式出版物面世，又有博客、播客等自媒体生产者持续输出。各类网络文献信息由于来源和表现形式不同，产生的信息没有相对统一的规范，且质量参差不齐。

(4) 网络文献分布广泛，具有无序性和隐蔽性，检索难度大。无论是传统的文字、图片、音乐信息，还是现有的视频、音频等信息，或是未来的增强现实、虚拟现实等表现方式，多种形式融合的网络文献分布在互联网的各个角落。其中有经过加工整理、结构规范、方便检索和获取的文献信息，但更多的文献信息处于无序状态，甚至具有一定的隐蔽性，检索或获取时需要多种技术手段进行文献处理才能进行统一管理和利用。

由于网络文献具有信息量大、生命周期短、形式多样等特点，所以与传统文献管理相比，网络文献必须根据其特点有针对性地进行研究管理。

### 3 网络文献的管理及其问题

李明杰教授<sup>[29]</sup>提出文献学研究应包括文献本体研究（如类型、产生等），文献整理研究（文献的保存、组织、挖掘等）和文献利用研究（文献的复制、流通等），且重在“文献整理和利用”。网络文献按照文献学路径研究的同时，更应关注其特质有针对性地进行管理。

#### 3.1 网络文献保存

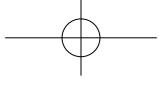
网络文献随网络技术的出现及广泛应用，在内容、形式、类型等方面不断扩充和延展。网络文献可以由规范的出版机构发行，如网络报纸、网络期刊，也可以由专业机构（政府、企业、组织）发布。网络文献可以是科研工作者为学术交流，遵循一定规范而完成的学术研究作品，也可以是普通网络用户对日常生活的简单记录。网络文献的产生几乎没有任何限制，只要不违背法律法规，具有表达的意志就可以在网络上生成属于自己的信息，成为“自媒体”。在网络文献发展初期，网络文献保存问题几乎不被关注，但随着网络文献数据量的增加，其中蕴含的社会和历史价值逐渐凸显，而网络文献的不稳定性可能导致依赖网络存在的人类社会记忆的消亡，因此，各个国家开始建立网络文献保存体系。

针对网络文献内容丰富、样式多样化的特点，互联网保存联盟资助并发布的 WARC 成为网络资源存档使用的文件格式。目前大多数的网络文献存档工作均采用此文件格式，如英国网络档案馆<sup>[30]</sup>、澳大利亚 Pandora 项目<sup>[31]</sup>、中国国家图书馆<sup>[32]</sup>、中国台湾大学网站典藏库项目<sup>[33]</sup>等。

网络文献保存根据文献信息覆盖范围可以分成 4 类：

##### (1) 国家（地区）范围

此类保存主要围绕国家或地区的重要历史事件、具有历史文化价值的网站、具有地方文化特色的网络资源开展，一般由国家级图书馆牵头进行信息采集和保存。目前此类网络文献保存项目数量最多，影响最广，如美国 Internet Archive 的 Archive-It 项目<sup>[34]</sup>、美国国会图书馆 MINERVA



项目<sup>[35]</sup>、大英图书馆和英国国家档案馆等机构发起的 UKWA 项目<sup>[30]</sup>、澳大利亚国家图书馆 Pandora 项目<sup>[31]</sup>、法国国家图书馆 BnF 项目、芬兰国家图书馆 FinnishWeb Archive 项目、瑞典国家图书馆 Kulturarw 项目等<sup>[36]</sup>。中国国家图书馆从 2003 年起开展网络文献采集和保存实验项目, 2007 年加入国际互联网保存联盟, 2014 年通过“网事典藏”项目联合全国的公共图书馆以全域采集和专题性采集完成政治、经济、文化发展的网站和专题资源保存<sup>[37-38]</sup>。

### (2) 机构范围

当网络文献影响范围缩小时, 就可以由指定机构或组织对其进行保存。机构范围的网络文献保存内容根据机构的特点、关注的内容、需长期保存的信息分类等确定其采集和存档的实现。以欧盟网络档案项目<sup>[39]</sup>为例, 该项目负责对托管的 europa.eu 域或其子域上的欧盟机构网站进行保存, 除了每 4 年进行的自动存档外, 还会对下线或发生变化的网站进行临时爬取保存, 也可以根据欧盟机构的特定要求进行信息采集和留存。英国议会网站档案<sup>[40]</sup>则用于采集和保存有关英国议会的网络文献信息, 包括 2009 年至今的英国议会网站和政务社交媒体信息。世界银行网络档案项目<sup>[41]</sup>保存已停用或大幅更新的世界银行网站的内容, 从而保留其历史和研究价值。

### (3) 社群范围

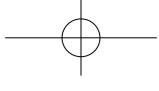
网络文献的网络传输特性让全球的共享交流更加便利, 越来越多的用户由于相同的身份、爱好、文化背景或信仰形成特定的社群。例如, 根茎组织为纽约市现代艺术博物馆附属机构<sup>[42]</sup>, 致力于创作、展示和保存用数字网络和工具制作的艺术品进而倡导自然的数字艺术和文化<sup>[43]</sup>。类似的社群还有 NYARC 网络资源<sup>[44]</sup>、荷兰声音与视觉研究所<sup>[45]</sup>等都将保存的网络文献类型定位为新时代多媒体类型展示的学术、教育、艺术等领域的作品。

### (4) 个人范围

随着移动网络的发展, 网络文献的生产者门槛降低, 普通大众参与到分享自身事实中, 成为自媒体的主力军。尽管普罗大众的个人记录具有极大的个性, 但无数个体的记忆汇聚能更加全面、细致地反映当代社会生产生活的原始面貌<sup>[46]</sup>。把有价值的个人网络信息加以归档保存, 发挥其社会记忆功能是个人范围网络文献保存的初衷。较为代表性的 Internet Archive (互联网档案馆)<sup>[47]</sup>致力于保存社交媒体上的信息, 尤其是即将关闭的平台, 如已明确不再为个人用户提供服务的 Google+ 社交网络平台。另外, 美国国会图书馆发起的社交媒体归档项目<sup>[27]</sup>, 中国国家图书馆与新浪微博在博文保存与利用方面建立合作关系<sup>[48]</sup>, 德国的 LiWA、ForgetIT 项目和英国的 ARCOMEM 项目的目标都是建立社交媒体的长期保存系统<sup>[49]</sup>, 但以上保存都是有选择性的归档, 更多个人范围的网络文献信息保存需要借助社交媒体网站提供的数据下载功能或第三方提供的信息备份工具完成本地保存, 这种保存方式并非符合网络文献要求的归档, 保存后的内容无法在原始数据丢失或更新的情况下, 恢复到网络上实现再次获取。在社会记忆理论的大背景下, 社交媒体的归档成为现在研究的热点内容之一。

## 3.2 网络文献利用

网络文献的呈现从原有的文字、图片和声音发展到今天的视频、音频甚至互动性极强的弹幕等多种形式。但从其本质上来看并没有脱离文献的基本利用方式即文献复制、流通、检索等。



### (1) 文献复制

网络文献的载体是网络,用户只需要通过“下载”就可以完成文献复制<sup>[50]</sup>。随着分布式融媒体技术在数字出版领域的深度应用,网络侵权现象愈发严重<sup>[51]</sup>。以我国网络文学付费阅读为例,网络盗版主体借助自媒体链接、社交分享、P2P 技术进行病毒式上传和下载数字版权内容,致使 2018 年我国 PC 端、移动端的网络文学付费阅读收入损失超过 70 亿元<sup>[52]</sup>。更遑论通过印刷、音频化或视频化等不同形式对原文献进行复刻或演绎,远远超过了静态复制的范畴。网络文献复制方式的不断变化和升级给知识产权保护带来巨大挑战。

### (2) 文献流通

网络文献的流通速度和范围取决于网络的延伸程度。网络文献的共享和交流是其重要特点,而现在的文献流通不同于传统文献流通可记录、可循迹的方式,用户可以通过转发、下载、超链接的方式完成信息流通,对于未授权的文献也可以采用现代化的拍照、摄像或第三方工具完成文献的复制和流通。文献作者或拥有者由于版权意识不足,可能忽略了网络文献的网络传播权的授权,也会造成文献无节制、无规范的流通。

### (3) 文献检索

网络文献检索依赖网络文献的规范性和编目情况。不同类型的文献检索方式有所不同。网络期刊、网络报纸、电子公告服务等类型由相关网站按照文献类型、文献服务对象、文献主题等提供搜索功能,便于用户快速检索得到需要的文献信息。网络数据库或预印本服务则提供相对专业和细致的检索条件帮助用户进行文献检索。其它文献在所属网站或平台上都提供了一般的查询检索功能,为用户提供快速定位文献功能。

文献呈现形式不同,检索方式也有区别。用户大多使用搜索引擎对网络文献进行检索,如 Google、WiseNut、百度等。现有的搜索引擎基本已实现全文搜索,可以自动爬取互联网上的网页信息,也可以通过网站主动提供网页信息获取网络文献内容。获取到的内容经过 PageRank 算法或分词计算筛选出要保存的信息,对每个词建立索引,并将索引保存到数据库中。用户提出检索需求后,搜索引擎将数据库中查询到的内容排序后提供给用户,用户可以通过超链接获得网络文献。对非文字文献类型的检索则需要通过该文献在网络呈现时提供的标签、关键词、类型等外部属性的文字性描述完成。

## 3.3 文献管理中存在的问题及解决思路

由于网络文献仍处于不断发展变化的过程中,现有的法律法规、技术不能解决发展过程中随时涌现的问题,故在网络文献管理的诸多环节如文献保存、复制、流通和检索中存在以下问题。

### (1) 知识产权问题

网络文献使用中,知识产权保护难度比较大。尽管采用先进的技术手段可以避免抄袭,但知识产权保护还应关注复制权、发行权、公共传播权等权利的侵权和保护<sup>[53]</sup>。网络文献尤其是社交媒体中的创作独特性的判断困难重重,在以作者生产内容的网站中,著作权的判别存在不同的标准<sup>[54]</sup>。

版权保护法律法规需要持续完善,2021年6月我国知识产权法中有关著作权、专利权等新规定的正式实施,将为网络文献的知识产权问题提供法律依据。文献的生产者、传播者、使用者



都应该提高版权保护意识, 只有全民知识产权意识提高, 才能在网络文献整个生命周期的各个环节都避免出现知识产权问题。同时要积极探索应用区块链技术实现版权的全流程可追溯管理<sup>[55]</sup>, 从技术层面为知识产权问题保驾护航。

### (2) 网络文献归档问题

由于网络文献存储载体的不稳定性、网络文献内容的持续扩充性以及现有归档的不完全性, 网络文献归档将面临更艰巨的任务。尽管各国均已在网络文献归档上投入大量的精力, 但效果仍不尽如人意。已归档内容更多集中于国家(地区)的重大事件, 归档主体仍集中于各国的政府机构、国家图书馆和档案馆<sup>[56]</sup>, 挖掘多种归档主体, 鼓励各种社会力量发挥文献归档作用, 才有可能实现对网络文献的全领域归档。

网络文献数量增速远大于归档能力的增长, 归档前必定要对网络文献进行评估和遴选, 从而确保归档文献的质量有保障, 发挥其社会和历史价值。这就要求在文献归档的规范和审核中发挥现代化技术的作用, 在不影响审核质量的前提下提高审核的自动化水平, 为大数据时代下网络文献的归档奠定基础。

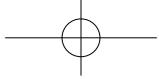
### (3) 网络文献检索效率问题

现有文献检索大多数仍然依赖于文字信息检索。目前, 图像类型的文献可以采用基于图像内容特征和图像形式特征进行检索, 例如拍照搜索功能, 但需要图像具有一定的规范性, 目前还未应用于大规模的网络图像文献的搜索中。直接用音频进行声音类文献搜索目前效果不理想, 其难点在于对声音事件进行检测, 根据特征提取音频并进行模式识别。视频文献可以看作由音频和动态图像组成, 故其搜索识别难度更大。综上所述, 由于文献类型和内容样式的差异化, 现有的文献检索效率仍然不高, 尤其是目前搜索引擎算法封装在商业公司提供的搜索服务背后, 商业公司由于经济、社会等各种利益关系会有针对性的对搜索结果进行干预, 导致文献搜索结果不能十分客观、准确, 也给文献搜索带来障碍。网络文献的搜索需要图像处理、音视频识别等领域的人工智能技术支持才可能达到用户期望的快速、准确、客观要求。这就要求文献研究人员更多地关注科学技术发展, 并将其应用到文献检索之中, 为文献检索提供更有效的技术手段。

### (4) 网络文献信息挖掘问题

大数据时代下的网络文献数量、内容和形式与传统文献相比都呈现出巨大的差异, 传统的线性平面表现形式被网络文献的动态立体的信息表现方式所替代, 如何将海量、碎片化、复杂的网络文献信息融合为系统化或创新化的知识, 成为网络文献信息挖掘的难点问题。

网络文献是科学技术发展的产物, 同时又推动了科技的变革。越来越多的研究者开始利用人工智能技术解决文献信息数据量巨大、载体多样、信息不完整等问题<sup>[57-58]</sup>。机器学习、知识图谱、计算机视觉、自然语言处理等前沿技术交叉融合, 在网络文献信息抽取、挖掘、可视化等方面发挥越来越重要的作用。在人工智能技术的加持下, 网络文献信息的跨时空、跨语种、跨媒介、跨领域的关联成为可能, 信息的网络化关联和流动转化使得“知识获取已非直接来自单个文本, 亦非来自文本本身”<sup>[18]</sup>, 知识将以更多元的形态和性能呈现给大众。

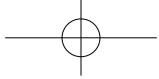


## 4 结语

网络文献作为新时期的产物,具有发展变化快、形式多样化、产生来源丰富等特点。笔者通过梳理和分析文献定义,发现文献研究实质由“知识”逐渐扩大为“信息”的趋势,而网络文献经过辨析符合现代文献定义,可以归属于“文献”范畴。基于当前网络文献产物的共性抽象出网络文献定义,融合传统文献类型和网络文献的呈现方式将网络文献分为九大类,最后从文献研究维度总结了网络文献整理和利用的现状及存在的问题。网络文献随现代化技术发展而产生,也必将随现代化技术的推进而持续演变,未来网络文献的形式、内涵和特点势必呈现动态变化的趋势,网络文献研究中存在的问题将成为文献信息学的研究内容,而现代信息技术和文献学研究深入融合,也将持续推动文献信息学的创新发展研究。

### 【参考文献】

- [1] 邵胜定. 说文献 [J]. 文献, 1985(04):163-172.
- [2] 董恩林. “文献”之我见——兼与单柳溪同志商榷 [J]. 文献, 1986(04):161-168.
- [3] 单柳溪. “文献”诠释 [J]. 文献, 1986(1):223.
- [4] 周启付. 什么是文献 [J]. 文献, 1987(02):190-198.
- [5] 洪湛侯. 中国文献学新编 [M]. 杭州: 杭州大学出版社, 1994.
- [6] 傅广荣. 文献定义“载体”辨析 [J]. 图书馆学研究, 1988(03):44-45.
- [7] 朱建亮. 论文献观 [J]. 图书情报工作, 1986(06):5-11.
- [8] 林申清. 现代文献的定义及其相关概念 [J]. 图书馆杂志, 1989(04):3-5.
- [9] 于鸣镝. 文献新说 [J]. 图书情报工作, 1995(03):27-30,53.
- [10] 陈界. “文献”定义的几个问题 [J]. 中华医学图书情报杂志, 2015,24(04):51-55.
- [11] 朱娜娜, 马海群, 张智钧. 文献的现象学定义 [J]. 图书情报知识, 2019(02):74-80.
- [12] 王雅玲. 试论文献载体三次变革对文献信息发展的影响 [J]. 闽西职业大学学报, 2003(02):63-65.
- [13] 王东波. 面向 21 世纪的网络文献传递业发展研究 [J]. 现代情报, 2001(05):71-72.
- [14] 相丽玲. 网络文献初探 [J]. 生产力研究, 2002(04):102-103.
- [15] 黄晓斌. 论网络文献的信息集成 [J]. 图书情报工作, 2003(01):20-26.
- [16] 刘佳. 网络文献的界定和评价 [J]. 情报科学, 2003(09):994-997.
- [17] 郭伟, 陈丽华. 网络文献概念的重新定义及范围界定 [J]. 吉林医药学院学报, 2009, 30(04):230-232.
- [18] 刘石, 李飞跃. 大数据技术与传统文献学的现代转型 [J]. 中国社会科学, 2021(02):63-81,205-206.
- [19] 2020 年世界互联网发展指数排名发布. 澎湃新闻 [EB/OL]. [2021-05-16]. [https://www.sohu.com/na/433772266\\_260616\(2021-02-23\)](https://www.sohu.com/na/433772266_260616(2021-02-23)).
- [20] 2020 全球数字报告. [EB/OL]. [2021-05-16]. <https://zhuanlan.zhihu.com/p/118217479>.
- [21] 网络期刊 [EB/OL]. [2021-05-16]. <https://baike.baidu.com/item/%E7%BD%91%E7%BB%9C%E6%9C%9F%E5%88%8A/3815534?fr=aladdin>.
- [22] 电子公告服务 [EB/OL]. [2021-05-16]. <https://baike.baidu.com/item/%E7%94%B5%E5%AD%90%E5%85%AC%E5%91%8A%E6%9C%8D%E5%8A%A1/5553364?fr=aladdin>.
- [23] 丁晓萍. 世界知名预印本数据库 arXiv [J]. 商丘师范学院学报, 2019,35(05):104-106.
- [24] 柳丽花, 王者乐. 国内 OA 站点的资源组织形式比较分析及启示 [J]. 中华医学图书情报杂志,



韩娜, 马海群, 张春春. 网络文献的内涵界定与管理 [ J ]. 文献与数据学报, 2021, 3 ( 2 ) : 035-046.

2013,22(03):14-18.

[ 25 ] 中国科技论文在线 [ EB/OL ]. [ 2021-05-16 ]. <http://www.paper.edu.cn/>.

[ 26 ] 社交媒体 [ EB/OL ]. [ 2021-05-16 ]. <https://baike.baidu.com/item/%E7%A4%BE%E4%BA%A4%E5%AA%92%E4%BD%93/1085698?fr=aladdin#9>.

[ 27 ] HENNINGER M, SCIFLEET P. How are the new documents of social networks shaping our cultural memory [ J ]. Journal of Documentation, 2016, 72(2).

[ 28 ] 王珊, 萨师焯. 数据库系统概率 (第 5 版) [ M ]. 北京: 高等教育出版社, 2014:4.

[ 29 ] 李明杰, 许小燕. 中国文献学学科体系的历史演变与现实重建 [ J ]. 图书情报知识, 2016(02):30-38.

[ 30 ] UK Web Archive. Save a UK website [ EB/OL ]. [ 2021-05-16 ]. <https://www.webarchive.org.uk/>.

[ 31 ] National Library of Australia. Pandora Overview [ EB/OL ]. [ 2021-05-16 ]. <http://pandora.nla.gov.au/overview.html>.

[ 32 ] 赵丹阳. 国家图书馆网络资源采集与保存平台的技术实现 [ J ]. 数字图书馆论坛, 2020(09):41-47.

[ 33 ] 台湾大学图书馆. 台湾大学网站典藏库 [ EB/OL ]. [ 2021-05-16 ]. <http://webarchive.lib.ntu.edu.tw/default.asp>.

[ 34 ] 王烁. 美国网页归档项目——Internet Archive 发展研究 [ J ]. 兰台世界, 2012(6):18-19.

[ 35 ] 时婉璐. 原生数字资源馆藏建设初探——以美国国会图书馆 MINERVA 项目为例 [ J ]. 图书馆杂志, 2013, 32(08):27-30, 26.

[ 36 ] HAKALA J. Archiving the web: European experiences [ J ]. Program, 2004, 38(3):176-183.

[ 37 ] 张炜, 敦文杰, 周笑盈. 国家数字图书馆网络资源保存的实践与探索 [ J ]. 数字图书馆论坛, 2017(06):32-38.

[ 38 ] 孙倩, 张炜. 我国图书馆开展网络信息保存的采选策略研究 [ J ]. 图书馆学研究, 2016(17):28-32.

[ 39 ] Publications office of the European Union. EUWeb archive [ EB/OL ]. [ 2021-05-16 ]. <https://op.europa.eu/en/web/web-tools/euwebarchive/>.

[ 40 ] UK Parliament Archive. UK Parliament Web Archive [ EB/OL ]. [ 2021-05-16 ]. <http://webarchive.parliament.uk/>.

[ 41 ] The Worldbank. The Worldbank Web Archives [ EB/OL ]. [ 2021-05-16 ]. <https://www.worldbank.org/en/webarchives>.

[ 42 ] 周文泓, 苏依纹, 代林序, 等. 进展与展望: 面向数字记忆资源建设的网络信息存档 [ J ]. 图书馆论坛, 2020, 40(09):42-52.

[ 43 ] Rhizome. About Rhizome [ EB/OL ]. [ 2021-05-16 ]. <https://rhizome.org/editorial/>.

[ 44 ] Nyarc. Nyarc Web Archiving [ EB/OL ]. [ 2021-05-16 ]. <http://www.nyarc.org/content/web-archiving>.

[ 45 ] The Netherlands Institute for Sound and Vision. Sound and Vision web archive [ EB/OL ]. [ 2021-02-18 ]. <https://www.beeldengeluid.nl/en/about>.

[ 46 ] 任洪展. 自媒体网络信息归档初探 [ J ]. 档案与建设, 2015(10):14-17.

[ 47 ] Internet Archive: 记录那些被遗忘的互联网 [ EB/OL ]. [ 2021-05-16 ]. <https://baijiahao.baidu.com/s?id=1650262387922652808>.

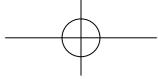
[ 48 ] 王露露. 中美图书馆社交媒体保存项目的比较与启示 [ J ]. 图书馆学研究, 2019(19):40-48, 39.

[ 49 ] 黄新平. 欧盟 FP7 社交媒体信息长期保存项目比较与借鉴 [ J ]. 图书馆学研究, 2019(17):2-9.

[ 50 ] 李艳. 图书馆文献复制与传播的版权问题研究 [ J ]. 河南图书馆学刊, 2020, 40(08):85-87.

[ 51 ] 中国信通院发布《2018 年中国网络版权保护年度报告》 [ EB/OL ]. [ 2021-05-16 ]. [http://www.sohu.com/a/311083122\\_120025397](http://www.sohu.com/a/311083122_120025397).

[ 52 ] 张冰清, 李琳. 基于区块链技术的数字版权利益平衡 [ J ]. 中国出版, 2019(11):22-25.



- [53] 曹作华. 网络文献版权保护若干问题探讨 [J]. 情报杂志, 2002(02):25-27.
- [54] 朱鸿军, 彭桂兵. 社交媒体空间的著作权: 西方的研究视角 [J]. 现代出版, 2020(04):48-55.
- [55] 穆向明. 基于区块链技术的数字版权保护新思路: 《2018年中国网络版权保护年度报告》评述 [J]. 出版广角, 2019(19):91-93.
- [56] 高晨翔, 黄新荣. 国外社交媒体文件归档的政策研究 [J]. 图书馆, 2017(07):6-11.
- [57] 王延飞, 赵柯然, 何芳. 重视智能技术 凝练情报智慧——情报、智能、智慧关系辨析 [J]. 情报理论与实践, 2016,39(2):1-4.
- [58] 曹文振, 赖纪瑶, 王延飞. 人工智能时代情报学发展走向之辨——对本体论、感知论、方法论、服务论的再思考 [J]. 情报学报, 2020(5):557-564.

## The Connotation Definition and Management of Network Document

HAN Na<sup>1,2</sup> MA Haiqun<sup>1</sup> ZHANG Chunchun<sup>1,3</sup>

- (1. School of Information Management, Heilongjiang University, Harbin 150080, China;  
2. School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150027, China;  
3. Heilongjiang Provincial Library, Harbin 150090, China)

**Abstract:** [ **Purpose/significance** ] In view of the rapid development of network document and the continuous extension of its connotation, the existing definition of network document has limitations, so the network document is redefined. [ **Method/process** ] Through sorting out the definition of literature and existing network document, combining the characteristics and connotation of network document in the era of big data, the research abstracts the definition of network document and researches network documents. [ **Result/conclusion** ] New definition, classifications and characteristics of network document are given. Based on the characteristics of network documents, the paper analyzes its current preservation and utilization status in the management, and proposes solutions to the problems of intellectual property protection, document archiving, retrieval efficiency and information mining. As a combination of literature and the development of information technology, the study of network document expands the scope of philology research and provides a new research perspective for document informatics.

**Keywords:** Network documents; Definition; Classification; Management

( 本文责编: 周 霞 )