DOI:10.13530/j.cnki.jlis.2022024

文档数据化:概念、框架与方法*

杨建梁　刘越男　祁天娇

摘　要　数据价值已经得到社会各界的高度认可。为进一步利用大数据、人工智能等技术释放数据的价值,文档数据化的概念被提出并日益受到重视,也成为图书情报与档案管理学科数字转型的新领域。经多学科概念与方法的综合和推演,本文对文档数据化的概念内涵、内容框架和关键方法展开系统研究。研究发现,文档数据化是面向文档的开发利用,将文档转变为机器可识别、可分析、可计算的数据的过程;智能技术允许机器参与到文档数据化的决策过程中,使得文档数据化呈现出人机协同、利用驱动、粒度细化、面向计算的特点。基于以上研究,本文提出文档数据化的任务框架,包含转录识别、描述增强、关联构建和矢量处理四项任务,呈现出结构化、语义化和智能化三个维度上面向机器的演进机制。对各项任务涉及的基础方法和关键方法进行梳理后可知,以深度学习、自然语言处理等技术为核心的文档数据化方法正在发挥越来越重要的作用。图6。表6。参考文献36。

关键词　文档　数据化　非结构化数据　结构化　量化

分类号　G255.51

Documents Datafication: Concept, Framework and Methods

YANG Jianliang, LIU Yuenan & QI Tianjiao

ABSTRACT

The value of data is becoming highly recognized. However, according to statistics, most data is unstructured documents that cannot be directly analyzed and calculated by machine, such as books, periodicals, documents, archives and so on, whose value is difficult to be fully released. In order to further utilize big data, artificial intelligence and other technologies to release the value of data, the concept of documents datafication has been put forward and focused. Datafication is becoming a new field of digital transformation in library science, information science and archival science. However, the concept of documents datafication is still vague and a systematic framework has not been formed yet.

In order to consolidate the conceptual foundation of our discipline and effectively promote the development of the theory and practice of documents datafication, the authors systematically conduct studies on the conceptual connotation, content framework and core methods of documents datafication through synthesis and deduction of multi-disciplinary concepts and related methods.

* 本文系中国博士后科学基金面上资助一等项目"基于深度学习与事件知识图谱的数字文书档案价值鉴定研究"(编号:2020M680029)的研究成果之一。(This article is an outcome of the key project "Value Appraisal of Digital Documentary Archives Based on Deep Learning and Event Knowledge Graph" (No. 2020M680029) supported by the China Postdoctoral Science Foundation.)

通信作者:刘越男,Email:liuyuenan@ruc.edu.cn,ORCID:0000-0002-5216-2111(Correspondence should be addressed to LIU Yuenan, Email:liuyuenan@ruc.edu.cn, ORCID:0000-0002-5216-2111)

Documents datafication is defined in this paper as a process which transforms documents into data that can be recognized, analyzed, and computed by machines for the purpose of development and utilization of information resource. Intelligent technologies allow machines to participate in the decision-making process of documents datafication, making the documents datafication present the characteristics of humachined-cooperation, utilization-driven, granularity-refined and computing-oriented. Based on the findings described above, the authors further put forward the task framework of documents datafication. It mainly includes four tasks: transcription recognition, description enhancement, linkage construction and vectorization processing. The four tasks present a machine-oriented evolution mechanism simultaneously on three dimensions, namely the structuration dimension, the semantic dimension and the intelligentized dimension. It is found that the methods centered on deep learning, natural language processing and other technologies for documents datafication are playing an increasingly important role after the fundamental and key methods involved in the four tasks of documents datafication are combed. 6 figs. 6 tabs. 36 refs.

KEY WORDS

Document. Datafication. Unstructured data. Structurization. Quantification.

0 引言

进入21世纪以来,世界各国频繁发布大数据相关政策,数据在当今世界的战略地位不言而喻。而可能带来颠覆式创新的智能技术的快速发展,更让数据成为人工智能技术与产业发展的重要基础设施。尽管数据的地位与价值已经得到了广泛认可,但是数据赋能的过程却受到资源形态的掣肘。有学者指出,大数据环境下非结构化数据在所有数据中约占95%,比例极高[1]。非结构化数据在计算机系统中一般以二进制对象的形式存在,难以被机器直接分析和计算,其价值也就难以被充分挖掘。图档博领域中大部分数字形态的材料都属于非结构化数据,比如数字图书、电子文件、数字档案、艺术品数字影像等,本文将这些非结构化数据统称为"文档"。要想应用智能技术开发文档资源,充分挖掘数据价值来提供多样化服务,首先需要将文档资源转换为机器可以处理、计算和分析的数据,即开展文档数据化。

对于文档数据化,图书馆学领域围绕资源描述、标引、关联等主题开展了丰富的研究,此类研究大多归结在"信息组织""知识组织"的范畴内[2],甚少将之与数据化关联。数字人文学者则打破藩篱,认为数据化是继数字化之后图档博数据处理的重要阶段[3],是形成智慧数据的基础[4]。在档案学界,档案数据化被认为是继档案数字化之后的下一个发展重点[5],其与扫描数据的识别[6]以及结构化数据[7]相关联,人工智能技术在其中的应用也被提及[8,9]。可见,文档数据化已经成为图书情报与档案管理学科数字转型的新领域。虽然"数据化"术语渐趋流行,但其概念在本学科领域仍旧模糊,未达成广泛共识,未形成系统性框架;各种方法应用令人眼花缭乱,缺乏系统性梳理。现有研究中频繁出现的"结构化""语义化""量化"等表述,也需要我们进一步去辨析。

针对现有研究的不足,本文对文档数据化的概念、框架和方法开展基础性研究,以便为文档数据化——图情档应予深耕细作的新领域夯实根基。首先,综合多学科"数据化"的概念,结合文档非结构化数据的特点,明晰文档数据化的内涵和特点;继而提出文档数据化的任务,对其特征进行梳理分析,形成文档数据化的内容框架;基于此框架,对各项数据化任务所涉及的关键方法进行梳理,以便为后续的方法研究明确定位,同时为相关实践提供参考。

1　文档数据化的概念

1.1　文档数据化的内涵

理解“文档数据化”,关键在于要对“数据化”形成一个清晰界定。不同学科背景的学者从不同角度出发对数据化内涵的理解存在差异。一般认为,“数据化”是作为一个独立的术语出现在各类文献中,与大数据概念的出现和发展密不可分[10]。2013 年,维克托·迈尔-舍恩伯格在其著作《大数据时代》中提出了数据化的概念,认为通过构建数据表的方式实现部分信息从计算机不可分析到可分析的“量化过程”即是数据化[11]。此后,哲学、新闻传播、图情档、数字人文等多个领域的学者就数据化的内涵开展讨论。通过对相关研究成果的梳理,归纳数据化内涵的构成要素,如表 1 所示。

表 1　数据化内涵的构成要素

构成要素	描　述	来源领域	提出者
结构化和量化	通过数据表对信息进行结构化,从计算机不可分析到可分析的量化过程。	计算科学	维克托·迈尔-舍恩伯格[11]
机器可识别和分析	数据化是将电子形态进一步转换为可识别的文本与可分析的数据。	数字人文	赵思渊[12]
数字比特结构化和颗粒化	数据化是将均匀、连续的数字比特结构化和颗粒化,形成标准化的、开放的、非线性的通用的数据对象。	新闻传播	姜浩[13]
以数据形式记录事物	以数据的形式记录事物的信息并分析数据之间的相关关系。	哲学	陈志伟[14]
机器可理解和表达	数据化的过程需要把内容变成机器可以理解和表达的数据,并藉由算法实现对其中蕴含知识的发现和挖掘。	图情档	冯惠玲[15]

尽管不同学科对于数据化的定义存在区别,但通过观察表 1 中的构成要素不难发现,数据化是一个让信息由机器不可分析到可分析的转化过程。数据化过程的核心包含结构化和量化两个子过程。所谓结构化,是指根据不同的应用需求对信息进行解构和定义的过程,结构化允许机器通过数据定义部分地理解并处理数据定义后的信息内容。所谓量化,是在结构化的基础上进一步对数据进行特征提取,使之能够被机器理解和计算。结构化与量化并不是先后衔接的,而是相辅相成的。数据化的结果是形成计算机可直接处理和分析的数据对象,从而使得人们能够借助计算方法、技术和资源来应对海量信息带来的挑战。结合上述讨论,本文认为数据化是一个从信息到数据的结构化和量化过程,其本质是将信息从物理空间或文件系统中的比特(Bits)转换为数据库系统中的字节(Bytes)。

根据对象,数据化可分为业务数据化和文档数据化两种类型,前者是指通过一定的技术手段将业务过程和结果以可量化计算的结构化数据形式记录下来,以支持业务分析,这是一个从无信息记录到可计算数据的过程;后者则是将非结构化数据(可能是电子或非电子形式的)转变为机器可计算数据,这是一个从信息记录到可计算数据的过程。业务数据化可能包含文档数据化。文档数据化的最终目的是文档资源的开发利用,刘永等[16]特别将文档数据化利用端的产品服务和部分中间过程纳入到文档数据化的内涵中。因此,本文将文档数据化内涵概括为“面向文档的开发利用,将文档转变为机器可识别、可分析、可计算的数据的过程”,具体的

利用需求决定了数据化过程中所使用的技术方法和数据化成果的形式。本文的后续讨论,皆以该内涵界定为基础。

1.2 文档数据化的特点

(1)人机协同

人机协同,即文档数据处理从以人为主体转换为人与机器共同作为决策主体。随着信息技术的快速发展,传统环境下以人工为主的工作方式难以应对大数据环境下的海量文档。面对文档资源开发利用的"质"与"量"的双重压力,迫切需要机器辅助人来开展有关工作,而机器处理的优势在于精确计算,这就需要将文档资源转变为细粒度的、知识化的数据形态,从而让机器可以对文档资源进行识别、分析和计算。随着人工智能技术的发展,机器已经能够从数据中学习,进而对客观世界中物理对象在数字世界中的数据映射进行判断,人脸识别、智能推荐等应用都可以被视作是机器学习了规则之后所作的判断。此类文档资源开发利用发生了实质上的主体转换,引入了机器作为新的主体。此外,在文档数据化的各个环节中,也存在着大量机器作为主体参与决策的部分,比如基于人工智能的手写文档转录识别、自动化文本序列标注等。因此,人机协同并不是指人操作计算机,而是指两者共同就数据化的过程、方法和结果进行决策,目前的工作模式是以机器的智慧来辅助人脑进行决策,未来是否会出现如智媒世界中"人机合一"[17]的情况尚无法判断。但毋庸置疑的是,人机协同贯穿了文档资源从数据化到开发利用的整个过程,这既是文档数据化的产生背景,也是文档数据化的重要特点。

(2)利用驱动

文档数据化不是独立存在的过程,而是围绕文档利用服务而展开,与文档利用服务需求有着紧密的关联[6]。利用需求决定着文档数据化的具体内容,文档数据化与文档开发及利用服务呈现出相互验证、需求衔接的关系。不同利用需求驱动的文档数据化过程、方法以及成果形式会存在差异。比如,对于档案全文检索的利用服务,对应的文档数据化成果应该包含全文数据库和索引数据库,所采用的数据化方法以数字化扫描件的转录识别为主。再如,对于文档的参考关联分析,以搜寻和分析共同引用了某个制度或标准的文档之间的关联为例,其文档数据化的成果至少包括全文数据库和相关语义描述及标注,所采用的数据化方法以转录识别和描述标注为主。不同利用服务对文档的分析开发提出了不同需求,而不同的分析开发过程又对文档资源的形式和状态提出了不同需求。本文将从利用需求到分析需求再到文档数据需求的过程称为利用需求的反向传播,图1展示了利用需求反向传播的基本逻辑。

(3)粒度细化

文档数据化并不是线性的过程,其对象呈现出层层递进的数据粒度细化的特点。文档数据化不能简单地等价为从文档扫描件到文档全文转录的单一过程,因为这种等价忽视了文本等内容数据并非严格的结构化数据,文本内容数据虽然可以满足基础的分析和计算,但是很难满足更高层次的利用需求。在利用需求的驱动下,文本内容数据往往需要进一步标注和描述,甚至转为结构化程度更高的数据对象。因此,文档数据化并不仅仅是转录内容,而是要面向文档开发利用,包含对文档数据形态进行持续性改变的各个环节。文档数据化本身可以视作数据粒度细化且数据再组织的过程,如图2所示。对于数字化文档扫描件,首先对其进行转录识别,形成扫描件的全文数据,该过程也被称为文本化;然后对文档内容进行标注和描述,形成结构化程度更高的描述数据;继而构建本体,揭示文档所含概念之间的相互关系;最后对其中的知识实体进行抽取、关联、对齐,形成知识图谱。在上述环节中,数据对象粒度不断细化,每个环节都以上一个环节为基础,每个环节都对内容数据进行了重新组织,各个环节堆叠组成了文档数据化。

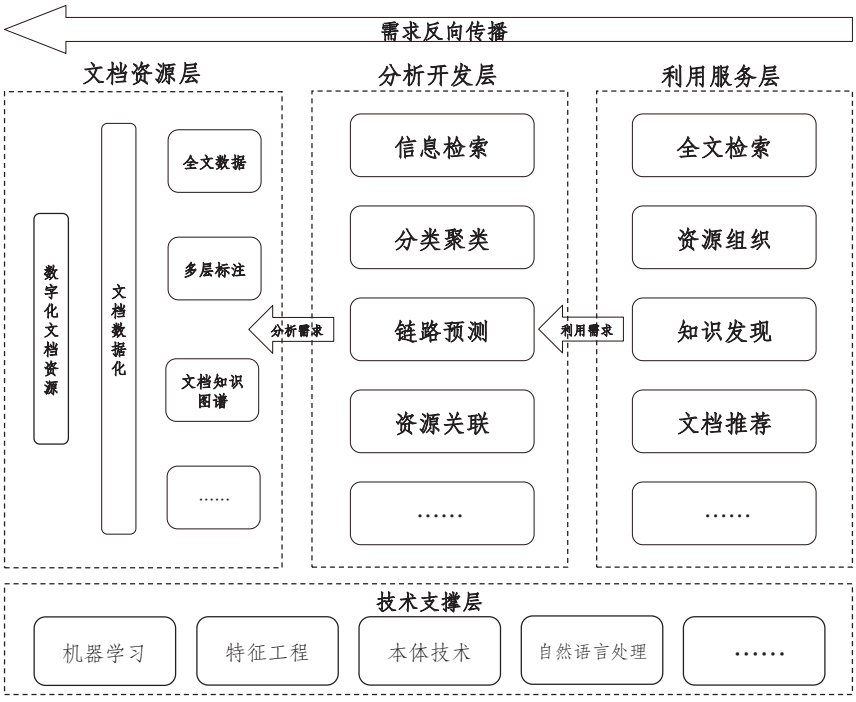

图 1　文档利用需求的反向传播

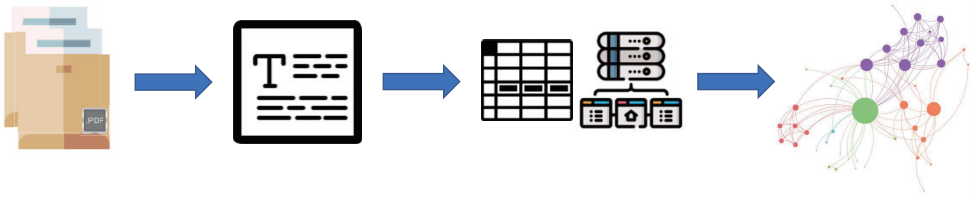

图 2　文档数据化过程中的数据粒度细化

(4)面向计算

所谓面向计算,是指文档数据化以机器可计算为导向。传统环境下的文档开发往往通过人工编目、编研等方式实现,人对于文档内容的阅读、梳理和分析的需求决定了文档资源的形态和组织是连续且易读的。而在新环境下,机器作为主体参与到文档开发利用中,机器识别和理解文档内容数据的过程,可以类比为传统环境下人对文档内容的阅读;机器计算可以类比为传统环境下人对文档内容的梳理和分析,如图 3 所示。而文档资源在进行数据化之前是面向人阅读的,并非面向机器计算。文档数据化是文档资源从面向人阅读到面向机器计算的转变。这种思想在数据化的内涵中也有所体现,量化作为数据化的核心内容之一,其要义是支撑数据可以被计算,从而支撑对数据的挖掘和分析。

1.3　数据化与数字化的关系

与数据化有着密切联系的另一个概念是“数字化”。数字化的概念先于数据化出现,尼葛洛庞蒂在其著作中将数字化定义为“将模拟信息转换成电脑可以处理的用 0 和 1 表示的二进制代码”[18]。该定义直观地表达了数字化的

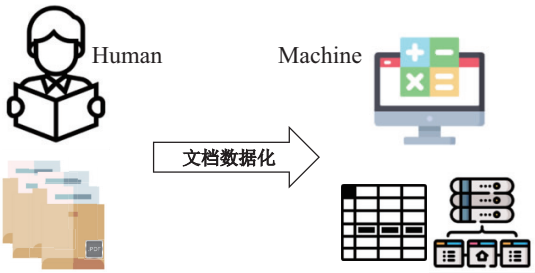

图 3　文档数据化实现从面向人阅读到面向机器计算的转变

本质是数据信号形式的转换。在图情档领域，尽管不同研究成果所给出的定义略有差别，但都将数字化与“把各类形式的信息输入到计算机系统并转化为二进制编码”进行深度关联[10,19,20]。

数据化以数字化为基础，是数字环境下数据处理领域继数字化之后新的发展重点，数字化与数据化既相互联系又相互区别。赵跃指出，其关键区别在于经过数字化和数据化处理后信息对象的形态不同[6]。这种理解一定程度上与本研究对数字化与数据化之间关系的描述相符。经过数字化的信息对象，其形态是二进制的比特，通过特定的软件或计算机算法排列组合成光学信号，主要面向人阅读和理解；经过数据化的信息对象，其形态是字节，是基于通用机器编码方式形成的数据，经过数据化的对象既面向人阅读和理解，也面向机器处理和分析。更直观地来看，对于载体为纸质的信息资源，数字化是将传统纸质载体上的信息进行扫描，结合著录元数据形成扫描件形式的数字资源。用户可以通过元数据对数字资源进行在线检索，并在线阅读扫描件；数据化则是进一步将扫描件形式的数字资源进行转录识别、描述增强、关联构建和矢量处理，将原本主要供人阅读的扫描件转变为数据库中的数据。

在某些语境下，数字化指代数字化技术或数字化服务。比如，柯平等指出图书馆的高质量发展需要以数字化技术赋能图书馆资源，实现图书馆信息资源的大数据化[21]。在这里，数字化对应的是人工智能、机器学习等数字化技术，这些数字化技术所解决的问题是实现信息资源的大数据化，使图书馆的信息资源变为“机器可读和可执行的数据集”[22]。而在数字图书馆领域，数字化也常与“服务”共用，常常指代数字图书馆所提供的数字化服务，比如在线阅览、参考咨询服务等[23,24]。在本研究的语境下，数字化或数据化均是指针对信息资源的处理过程，而非技术或服务，两者的主要区别如表 2 所示。

表 2　数据化与数字化的区别

	数据化	数字化
目的	将各类形式的信息输入到数据库系统中，面向机器分析计算	将各类形式的信息输入到计算机系统中，面向人阅读
主要任务	转录识别、描述增强、关联构建、矢量处理	扫描著录
资源状态	存储在数据库中的数据	存储在计算机中的数字资源

总的来说，数据化与数字化对信息资源的处理过程和结果有着明显的差别，前者强调将资源处理为字节，后者则是将资源处理为比特。相比于数字化，数据化更加面向机器理解、处理和计算，与大数据、人工智能等信息技术的联系更加紧密。但是，从数字化到数据化纵深发展

的过程中,数据处理侧重点的变化不是一蹴而就的,而是循序渐进的。数字化加工包含以光学字符识别(Optical Character Recognition,OCR)为代表的识别技术的应用,以及信息资源的元数据著录;到了数据化阶段,则进一步涌现出以神经网络、语言模型为代表的识别技术,以及对信息资源内知识单元的语义描述。所以我们不应割裂地看待数字化和数据化,两者的任务存在衔接和交叉,但数据化的任务更为丰富和深入,并且具有强烈的面向机器计算的特性。

2　文档数据化的内容框架

2.1　文档数据化的任务框架

根据文档数据化的内涵、特点以及文档开发利用的一般需求特征,本文构建了文档数据化的任务框架,如图4所示。

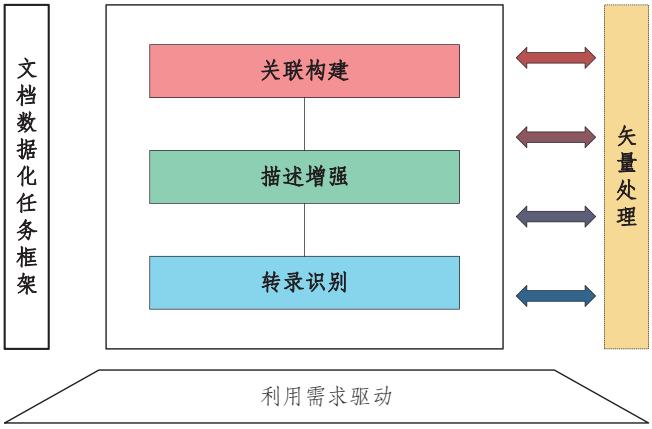

图4　文档数据化的任务框架

在该框架中,文档数据化受利用需求驱动,以数字化为基础。需要说明的是,对于经数字化加工转换的文档的数据化,其数据化和数字化工作存在一定的交集。在数字化阶段,可能开展了扫描件的小规模识别和手工为主的转录、文档级别的元数据著录,但这样的处理主要面向人的阅读和分析。最终面向机器计算的文档数据化被划分为四类任务,分别是转录识别、描述增强、关联构建和矢量处理。其中前三个数据化任务对应的是数据化的结构化过程,呈现层层递进的形式,而矢量处理对应数据化的量化过程,可直接应用于转录识别、描述增强、关联构建任一任务的成果。各任务模块的目标、产出和所处理的数据粒度如表3所示。需要说明的是,本研究所提出的任务框架并不涉及对具体任务成熟度或完成度的考量,文档数据化过程本身应根据利用需求有选择地开展。

表3　文档数据化的任务、目标、产出和数据粒度

任　务	目　标	产　出	数据粒度
转录识别	内容可操作	内容数据	文档级
描述增强	数据可理解	数据标注	文件、句子、词等多层级标注
关联构建	知识可获取	知识关联	实体级
矢量处理	机器可计算	特征向量	向量级

(1)转录识别

转录识别用于解决机器难以操作文档内容的问题。转录识别是文档数据化过程中重要且基础性的工作,其过程是将文档的内容数据和部分元数据有序存储在数据库字段或键中。以版式文件或扫描件为例,通过转录识别操作可以形成对应的全文数据库,从而使得管理者能够应用相关技术对文档内容和元数据进行全文检索、统计分析和可视化等工作。转录识别的目标可以被归纳为“内容可操作”。需要说明的是,对于以文本内容为主或者数据化成果以文本数据为主的文档,由于文本数据自身带有从篇章到语句再到词的层次性,其转录识别工作可以分为初级阶段和高级阶段。初级阶段是指将全部文本数据放到一个或几个字段中存储;高级阶段是指通过 XML 等标记语言对文本数据进行半结构化处理,将文本数据转化成文档树,再将文档树存储到数据库中,以区分文档内容的各个组成部分。

(2)描述增强

描述增强用于解决文档缺乏描述和标注,难以被机器理解的问题。描述增强是通过人工或机器对文档的内容数据和元数据进行标注。按照描述对象,标注层级可以包括多个层次,以文本文档为例,标注层次包括文件级、语句级、词级等;再以图像为例,标注层次包括图像级、对象级、像素级等。其中,信息资源元数据可以被认为是一种文件级的标注。经过描述增强,管理者可以通过机器对文档进行再组织。描述增强层所形成的成果一般包括富语义描述数据库和标注数据集,富语义描述数据库中包含丰富的描述数据,标注数据集中包含标注字段和记录。描述增强的目标可以被归纳为“数据可理解”。

(3)关联构建

关联构建用于解决文档资源的知识粒度大,缺乏细粒度知识表达的问题。关联构建是通过人工或机器对文档内容和元数据进行知识建模、信息抽取、关联揭示、知识融合等工作,从而通过文档数据之间的关联促进知识发现。其目的是将文档中蕴涵的相关知识通过知识图谱等方式表达出来,实现知识显性化、自动推理、知识发现以及智能审计、智能校验、智能风控等更高层次的智能化应用。通过关联构建,文档被转变为图结构的知识图谱,存储在图数据库中,或者变为文档资源及其描述信息的关联数据。关联构建的目标可以被归纳为“知识可获取”。

(4)矢量处理

矢量处理用于解决内容数据无法被机器计算分析的问题。矢量处理是机器自动化分析和处理文档资源的基础,通过相关算法对文档数据化后形成的各类结构的数据进行特征工程或表示学习,形成文件级、语句级、词级的向量表示,将文档、文档组成元素和知识实体映射到向量空间中。矢量处理是支持文档智能化利用服务的关键环节,能使机器实现自动化和智能化的应用,如主题聚类、多维分类、序列分析等。此外,矢量处理也是机器智能支撑文档数据化工作的重要内容,比如构建知识图谱时,往往需要通过矢量处理和人工智能技术,自动地抽取知识实体和关系。矢量处理的目标可以被归纳为“机器可计算”。

2.2 文档数据化的三维演进

文档数据化任务框架中的各个任务面向文档资源开发利用的不同问题,通过层层递进,逐个解决问题。从资源视角看,可以从结构化、语义化和智能化三个维度进一步理解文档数据化的任务,文档数据化不同任务的三维演进如图 5 所示。需要注意的是,无论是结构化、语义化还是智能化,都是主要面向机器处理的。

(1)结构化程度提高

文档数据化的关键任务之一是将文档资源转变为结构化的数据。受文档资源特性和利用需求的影响,文档数据化的各个任务呈现结构化程度逐步提高的趋势。通过转录识别,文档资源由机器不可直接操作的“块状数据”转变为

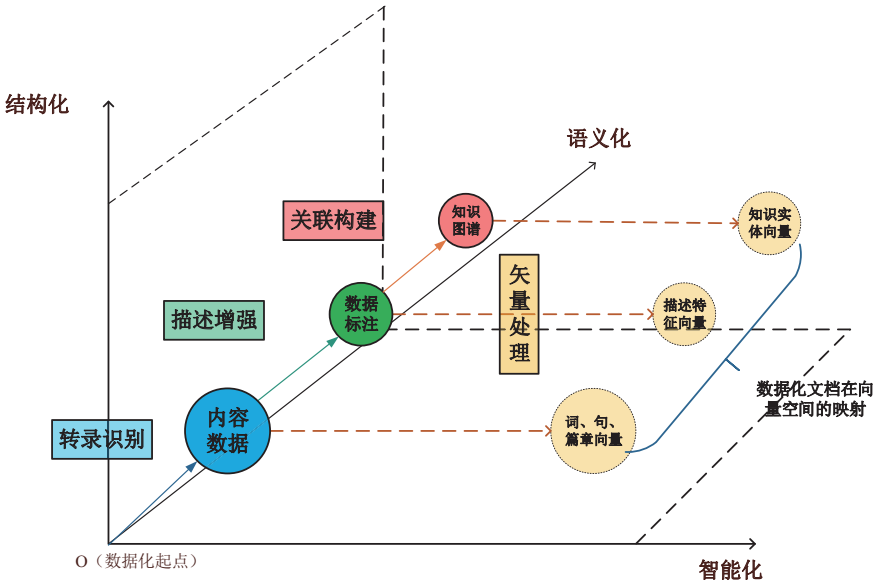

图 5　文档数据化的三维演进示意

可直接操作的内容数据，其成果往往以全文数据库的形式呈现。经过转录识别的文档资源可以被认为是一种初级的结构化数据。通过描述增强，已经被识别转录的文档内容数据获得了更丰富的人工或机器的描述和标注，这些描述和标注信息往往以结构化的数据来表示，由此文档资源被进一步结构化。通过关联构建，文档资源中的知识实体及关系被抽取出来，形成了知识图谱；文档资源通过其描述数据进行关联，形成“资源—实体”的关联网络。在此过程中，数据粒度进一步降低，结构化达到了最高程度。严格来说，矢量处理并没有直接参与提高文档资源的数据化程度，但矢量处理是机器支撑结构化的重要基础，能够间接地提高文档资源的结构化程度。

（2）语义化水平提升

随着文档数据化任务层次的提升，文档资源的语义化水平也在同步提升。这里的“语义化”，是指文档资源语义被精准揭示的程度和丰裕程度。在转录识别之前，文档资源的描述主要依靠自带的元数据。经过转录识别，文档资源的内容能够被机器所处理，进而提高了机器对文档资源的可读性。描述增强是提高文档资源语义化水平的关键过程，经过描述增强，管理人员可以为文档资源赋予不同层次粒度的描述和标注，这进一步增强了机器对文档资源的理解。关联构建通过揭示文档资源及其描述数据中的关联，从而发现新知识，这是更高层次的语义化体现。矢量处理采取了一种更加直接的方式，将文档映射到向量空间中，让机器可以直接计算和分析。从机器可读的视角来理解，矢量处理层也提高了文档资源的语义化水平。

（3）智能化能力增强

文档数据化各个任务层次支撑的利用服务的智能化程度呈现出增强的态势。经过转录识别，文档资源被转录为机器可操作的内容，为全文检索、主题聚类等应用提供了基础。经过描述增强，文档资源的描述和标注进一步丰富，可以由机器实现特定标准的分类、实体识别等更加智能的应用。经过关联构建，文档资源以知识图谱、关联数据等形式呈现，机器能够进一步开展自动推理、自然语言检索等高级别的智能化应用。矢量处理是智能化应用的关键步骤，也是将文档资源从面向人阅读转到面向机器计

算的重要步骤。经过转录识别、描述增强、关联构建后所形成的数据化产出,均可采用不同的技术方法进行矢量处理。比如,对于转录识别后形成的文本内容数据,可以按需处理为篇章向量、句子向量、词向量或字向量;对于描述增强后形成的数据标注,可以按需处理为标签向量、特征值向量等;对于关联构建后形成的知识图谱,可以处理为实体表示向量、图表示向量等。

3 文档数据化的方法体系

根据上文所提出的文档数据化任务框架,可构建出文档数据化的方法体系,如图 6 所示。

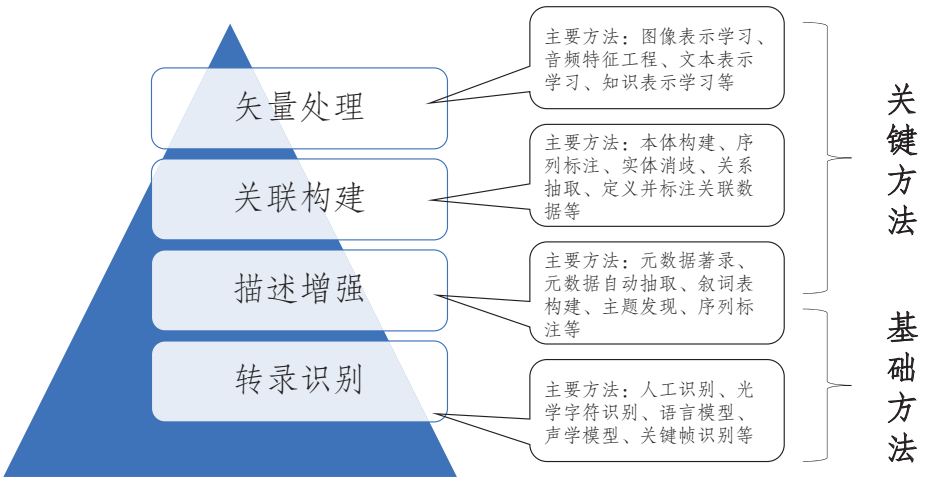

图 6 文档数据化的方法体系

根据文档数据化任务框架中各类任务的目标、产出和数据粒度,本文进一步从任务对象、实施主体两个方面对数据化的基础方法进行分析归纳。其中,任务对象是指文档数据化的具体对象类别,比如转录识别的对象一般包括印刷文档扫描件、手写文档扫描件以及音视频文档等。实施主体是指数据化的主要执行者,主要包括人和机器两类,其中机器作为实施主体是指机器参与了数据化过程中的决策过程,比如字符识别、描述标注等。一般来说,数据化的任务主体受到文档资源规模和利用需求的双重影响。资源规模越大,机器的参与度越高;利用服务的智能要求越高,机器的参与度也越高。在实际场景中,人和机器往往联合开展数据化工作。

3.1 面向转录识别和描述增强的数据化基础方法

转录识别作为文档数据化任务框架中的基础任务,任务对象主要包括印刷文档扫描件、手写文档扫描件、音视频文档和其他类别(如工程图纸档案),前置任务是以扫描著录为核心的数

根据方法的应用目的,本文将文档数据化方法分为基础方法和关键方法。基础方法旨在将文档资源转换为计算机系统中的二进制编码,并加以简单描述,其中代表性的方法包括扫描件 OCR 识别、元数据著录等,这部分方法和数字化方法存在交集。关键方法是面向机器可直接对文档资源计算和分析的数据处理方法,这些方法的应用可以使文档资源的结构化程度提高、语义化水平提升且智能化能力增强,其中代表性的方法包括实体及关系抽取、文本表示学习、主题发现等。本文在概要阐述文档数据化基础方法的同时,重点梳理数据化的关键方法。

字化任务。当转录识别的主要实施主体是人时,对于各类文档资源可以采取人工转录识别的方式,即通过人工阅读和分析,将文档资源识别转录为文本内容。当转录识别的主要实施主体是机器时,对于不同类型的文档资源对象,所采取的技术方法也有所差别。对于文档扫描件,其转录识别的关键方法主要包括 OCR[25]、手写识别和语言模型[9]。对于音视频文档,其转录识别的关键方法主要包括声学模型、语言模型、关键帧识别、图像识别等。结合视频分析的相关研究[26],视频文档的数据化分为对音频的数据化和对图像的数据化。对于图像部分,所采用的关键方法主要包括关键帧识别和图像识别。

描述增强的对象按资源粒度可分为文档集合(比如图书中的作品集、系列图书等,档案中的全宗、类目、案卷等)、文档件(单件)、语句和字词(主要针对的是文本数据)等。在数字化任务中,对文档件的元数据著录属于描述增强的方法。对于不同数据粒度的文档对象,其描述增强的关键方法也有所不同,数据粒度越细,机器能够参与的工作越丰富,数据粒度越粗,描述增强对人工的依赖越大。表 4 以文本类文档的数据化为例列出了描述增强的关键方法。

表 4　描述增强的关键方法

对象类别	实施主体	关键方法
文档集合	人	叙词表构建、本体建模、元数据著录等
	机器	主题模型等
文档件	人	元数据著录等
	机器	元数据抽取、分类标注、聚类标注、主题发现等
语句	人	—
	机器	语义角色标注、分类标注、聚类标注等
字词	人	叙词发现等
	机器	分词、词性标注、实体抽取、词义消歧、叙词发现与映射等

对于文档集合,主要需要人来开展描述增强工作,所涉及的关键方法包括叙词表构建、本体建模和元数据著录等。而机器可以进行主题发现和主题标注的工作,常见的技术方法有主题模型等。对于文档件的描述增强,人主要开展元数据著录等工作,机器参与的关键方法包括元数据抽取、分类标注、聚类标注和主题发现等。对于语句的描述增强,人不参与此工作,机器参与的关键方法包括语义角色标注、分类标注和聚类标注等。对于字词的描述增强,人主要开展叙词发现等工作,机器实施的关键方法包括分词、词性标注、实体抽取、词义消歧、叙词发现与映射等,所涉及的技术有隐马尔可夫模型、条件随机场、端到端的循环神经网络等。应用研究方面,倪渊通过上述方法,对电子病历中的字词进行了描述和标注,大大提高了电子病历的数据可理解性[27]。

3.2　面向关联构建的数据化关键方法

关联构建所涉及的数据对象主要包括领域本体、命名实体、实体关系和资源关系,这四类对象是基于文档资源构建知识图谱和“资源—实体”关联网络时所重点关注的数据对象。由于知识图谱和“资源—实体”关联网络的构建对数据处理的效率和实施主体的能力有较高的要求,现阶段知识图谱和“资源—实体”关联网络的构建往往是人与机器合作完成的(见表 5)。

表 5 关联构建的主要方法

对象类别	实施主体	关键方法
领域本体	人	概念术语分析、本体构建等
	机器	—
命名实体	人	人工识别并标注实体等
	机器	序列标注、实体消歧等
实体关系	人	人工识别并获取关系等
	机器	基于规则的模式识别、依赖路径识别、序列标注等
资源关系	人	定义关联数据模式、人工析出并整理实体与资源之间的关联关系等
	机器	自动标注并链接实体与资源等

领域本体构建是文档知识图谱构建的基础性工作,是定义概念及其关系的过程。这个过程需要由领域专家和信息资源管理专家联合开展,所采用的关键方法包括概念术语分析、本体构建等。机器基本不参与领域本体构建的相关决策。

命名实体和实体关系是指文档知识图谱中的知识实体及其关系,需要在文档内容中抽取并进行清理消歧。人工开展知识实体和实体关系的抽取工作时,需要对文档内容和领域本体具备一定的知识背景。对大规模文档内容采用人工方式进行知识实体及关系的抽取和消歧时,往往还需要通过众包或外包等方式由多人联合开展,如 Ellul 等研究提出,通过招募志愿者来协助公证档案知识图谱的知识抽取工作[28]。机器开展该项工作时,所采用的关键方法主要包括序列标注、实体消歧、基于规则的模式识别、基于依赖路径的识别等方法。序列标注一般采用条件随机场、循环神经网络等算法模型实现。实体消歧一般采用上下文相似度匹配等算法模型实现。基于规则的模式识别和基于依赖路径的识别是指按预定义的句法规则或实体关联规则抽取实体关系,一般采用深度神经网络等算法模型实现。现阶段已经有研究开展了基于文档资源的知识图谱构建,如面向科研档案管理的知识图谱系统构建方案[29]、面向公共危机事件的知识图谱[30]和中国历代存世典籍知识图谱[31]等。

资源关系是指文档中知识实体与资源之间的关联关系。人工开展资源关系的构建工作时,所采用的方法主要包括定义关联数据模式、人工析出并整理实体与资源之间的关联关系等。机器开展该项工作时,所采用的方法主要包括自动标注并链接实体与资源等。

3.3 面向矢量处理的数据化关键方法

矢量处理是机器参与数据化工作的基础,该任务贯穿于从转录识别到关联构建的各项数据化任务中。矢量处理的对象主要包括文档扫描件、音视频文档、转录识别后的文档件、转录识别后的文档语句、转录识别后的文档字词、知识图谱中的命名实体(见表 6)。由于矢量处理的主要目的是建立文档资源到向量空间的映射关系,实现机器可计算,因此主要实施主体是机器。

对于尚未转录识别的文档扫描件和音视频文档,机器可以采用图像表示学习、音频特征工程等关键方法来实现资源特征的矢量化,从而使机器参与到上述资源转录识别的过程中。

对于转录识别后的文档件、文档语句和文档字词的矢量处理,关键方法主要包括基于内容的表示学习、基于词袋的特征工程、字词表示

表 6　矢量处理的主要方法

对象类别	实施主体	关键方法
文档扫描件	机器	图像表示学习
音视频文档		音频特征工程、图像表示学习
转录识别后的文档件		基于内容的表示学习、基于词袋的特征工程（词袋构建可能需要人工参与）
转录识别后的文档语句		
转录识别后的文档字词		字词表示学习
知识图谱中的命名实体		知识表示学习、图表示学习

学习等。基于内容的表示学习主要通过深度学习模型或迁移训练的语言模型将文档内容表示为特征向量,常用的深度学习模型包括长短期记忆网络、BERT 等。基于词袋的特征工程主要通过构建特征词袋(词袋构建可能需要人工参与),进一步通过特征权重算法形成文档件和文档语句的特征向量,常用的特征权重算法主要包括 TF-IDF、互信息等。基于字词的表示学习主要基于字词的上下文关系,通过预训练或迁移训练模型获得词的分布式表示,常用的算法模型包括 Word2Vec、BERT 等。

对于知识图谱中命名实体的矢量处理,关键方法主要是知识表示学习和图表示学习。知识表示学习侧重于知识元组之间的推断关系,将知识图谱中的三元组近似为代数关系,基于向量平移不变性计算知识实体的分布式表示,常用的算法模型包括 TransE 等。图表示学习侧重于知识图谱的网络关系,利用节点之间的关联关系计算知识实体的分布式表示,常用的算法模型包括图卷积网络等。

4　结论与展望

本文以文档数据化的研究背景为起点,探讨了文档数据化的概念、框架和关键方法。首先基于多学科的相关文献,经过综合归纳推演出文档数据化的概念,即“面向文档的开发利用,将文档转变为机器可识别、可分析、可计算的数据的过程”,并总结出文档数据化呈现人机协同、利用驱动、粒度细化、面向计算等特点。其次,根据文档利用需求的反向传播路径,提出了文档数据化的任务框架,包含转录识别、描述增强、关联构建和矢量处理四项任务,随着任务的推进,呈现出结构化、语义化和智能化三个维度上的面向机器的演进机制。最后结合文档数据化各项任务的对象和实施主体,归纳梳理了各项数据化任务涉及的主要方法。

在新文科建设的背景下,冯惠玲[32]、马费成[33]、孙建军[34]、柯平[35]等众多专家指出图情档学科的发展要扩大学科空间、关注社会需求、拥抱智能技术、重视交叉融合。在社会高度认可数据价值,且数据概念延展至“任何以电子或者非电子形式对信息的记录”[36]的情况下,图情档这个从传统数据管理领域生长出来的学科,应该切实围绕当代社会开发数据生产要素价值的巨大需求,积极发展与之匹配的理论、方法和实践。可供智能分析的数据并不能凭空而来,对于占数据总量较大比例的非结构化文档而言,只有经历数据化过程,才能实现机器可操作、可理解和可计算,这是诸多机构所面临的现实问题,也是释放数据价值的关键环节,具有鲜明的社会需求属性。图情档在数据化问题上已经具备一定的研究和实践基础,文档数据化还可能将部分非图情档学科传统研究对象的数据资源纳入到研究范畴之内,从而开拓学科疆域。文档数据化正在成为图情档融入数字人文、人

工智能等学科的关键节点。因此,对于文档数据化这种既立足于社会需求,又具备跨学科属性的问题,需要图情档学科联合发声,以坚实的概念和稳健的框架为基础,实现跨学科的融合创新,积极占领新高地,提升图情档学科的社会影响力。

参考文献

[1] Gandomi A, Haider M. Beyond the hype: big data concepts, methods and analytics[J]. International Journal of Information Management,2015,35(2):137-144.

[2] 贾君枝. 面向数据网络的信息组织演变发展[J]. 中国图书馆学报,2019,45(5):51-60. (Jia J Z. Development of information organization for the data web[J]. Journal of Library Science in China,2019,45(5):51-60.)

[3] 曾蕾,谭旭. 数据的语义增强——解读图档博数字人文的新动向[J]. 数字人文研究,2021,1(1):65-86. (Zeng L,Tan X. Semantic enrichment of data: interpreting the new trend of LAM data in supporting digital humanities[J]. Digital Humanities Research,2021,1(1):65-86.)

[4] 曾蕾,王晓光,范炜. 图档博领域的智慧数据及其在数字人文研究中的角色[J]. 中国图书馆学报,2018,44(1):17-34. (Zeng L,Wang X G,Fan W. Smart data from libraries, archives and museums and its role in the digital humanity researches[J]. Journal of Library Science in China,2018,44(1):17-34.)

[5] 钱毅. 技术变迁环境下档案对象管理空间演化初探[J]. 档案学通讯,2018(2):10-14. (Qian Y. A preliminary study on the evolution of archival object management space under the environment of technological change [J]. Archives Science Bulletin,2018(2):10-14.)

[6] 赵跃. 大数据时代档案数据化的前景展望:意义与困境[J]. 档案学研究,2019(5):52-60. (Zhao Y. Prospects for archives datafication in big data era: significance and dilemma[J]. Archives Science Study,2019(5):52-60.)

[7] 赵生辉,胡莹. 档案文本结构化:概念、原理与路径[J]. 浙江档案,2019(12):23-25. (Zhao S H,Hu Y. Archival text structuring: conception, principles and approaches[J]. Zhejiang Archives,2019(12):23-25.)

[8] 杨建梁,刘越男. 机器学习在档案管理中的应用:进展与挑战[J]. 档案学通讯,2019(6):48-56. (Yang J L,Liu Y N. The application of machine learning in archives management: progress and challenges[J]. Archives Science Bulletin,2019(6):48-56.)

[9] Firmani D,Maiorino M,Merialdo P,et al. Towards knowledge discovery from the Vatican secret archives-In Codice Ratio-episode 1: machine transcription of the manuscripts[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York,USA,2018:263-272.

[10] 于英香. 档案大数据研究热的冷思考[J]. 档案学通讯,2015(2):4-8. (Yu Y X. Sober thinking about the craze for research on the archival big data[J]. Archives Science Bulletin,2015(2):4-8.)

[11] 维克托·迈尔-舍恩伯格,肯尼思·库克耶. 大数据时代:生活、工作与思维的大变革[M]. 盛杨燕,周涛,译. 浙江:浙江人民出版社,2013:104. (Mayer-Schönberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think[M]. Sheng Y Y,Zhou T, trans. Zhejiang: Zhejiang People's Publishing House, 2013:104.)

[12] 赵思渊. 地方历史文献的数字化、数据化与文本挖掘:以《中国地方历史文献数据库》为例[J]. 清史研究,2016(4):26-35. (Zhao S Y. The digitization of local historical archives, creation of metadata, and datamining: the example of *The Chinese Local History Archive*[J]. The Qing History Journal,2016(4):26-35.)

[13] 姜浩. 数据化:由内而外的智能[M]. 北京:中国传媒大学出版社,2017:17. (Jiang H. Datafication: an inside-out intelligence[M]. Beijing: Communication University of China Press,2017:17.)

[14] 陈志伟. 大数据方法论的新特征及其哲学反思[J]. 湖南师范大学社会科学学报,2020,49(1):24-31. (Chen Z W. New features of big data methodology and a philosophical reflection on them[J]. Journal of Social Science of Hunan Normal University,2020,49(1):24-31.)

[15] 冯惠玲. 融入数据管理 做电子文件管理追风人[J]. 北京档案,2020(12):6-7. (Feng H L. Integrating into data management to be into the wind trailer of electronic records[J]. Beijing Archives,2020(12):6-7.)

[16] 刘永,庞宇飞. 档案数据化之原生数据源全链式管理分析[J]. 档案管理,2018(5):11-18. (Liu Y, Pang Y F. Analysis for whole chain management of native data sources in archive digitization[J]. Archives Management,2018(5):11-18.)

[17] 彭兰. 人与机器,智媒化时代的主角之争[N/OL]. 社会科学报,2017-03-30(05)[2021-03-04]. https://www.sohu.com/a/131689568_550962. (Peng L. Human and machine, the fight for protagonists in the age of intelligent media[N/OL]. Newspaper of Social Science, 2017-03-30(05)[2021-03-04]. https://www.sohu.com/a/131689568_550962.)

[18] 尼古拉·尼葛洛庞帝. 数字化生存[M]. 胡泳,范海燕,译. 海口:海南出版社,1997:2. (Negroponte N. Being digital[M]. Hu Y, Fan H Y, trans. Haikou: Hainan Publishing House, 1997:2.)

[19] 王知津,潘永超. 数字图书馆合理使用问题研究[J]. 图书馆学研究,2009(1):21-24,59. (Wang Z J, Pan Y C. Research on the reasonable use of digital library[J]. Research on Library Science,2009(1):21-24,59.)

[20] 顾朝晖,朱伟铃,孙红卫. 数字图书馆信息自由权和知识产权的冲突[J]. 现代情报,2008(9):73-75. (Gu Z H, Zhu W L, Sun H W. The conflict of information freedom and property rights in digital library[J]. Journal of Modern Information, 2008(9):73-75.)

[21] 柯平,彭亮. 图书馆高质量发展的赋能机制[J]. 中国图书馆学报,2021,47(4):48-60. (Ke P, Peng L. Empowerment mechanism of library high-quality development[J]. Journal of Library Science in China,2021,47(4):48-60.)

[22] Lippincott S. Advancing digital scholarship[EB/OL]. [2021-08-21]. https://www.arl.org/wp-content/uploads/2020/07/2020.07.06-emerging-technologies-advancing-digital-scholarship.pdf.

[23] 徐健晖. 国外高校图书馆按需数字化服务实践与启示[J]. 国家图书馆学刊,2018,27(2):68-74. (Xu J H. Practice and enlightenment of digitization services on demand in foreign university libraries[J]. Journal of the National Library of China,2018,27(2):68-74.)

[24] 夏梦蝶,薛岳,候梦洁. 科研院所图书馆数字化服务探究[J]. 图书情报工作,2017,61(S1):70-72,84. (Xia M D,Xue Y,Hou M J. Research on digital service of research institute library[J]. Library and Information Service,2017,61(S1):70-72,84.)

[25] 国家档案局. 纸质档案数字复制件光学字符识别(OCR)工作规范: DA/T77—2019[S]. 北京:国家档案局, 2020: 3-9. (National Archives Administration of China. Specification for Optical Character Recognition (OCR) of digital copies of paper-based records: DA/T77—2019[S]. Beijing: National Archives Administration of China, 2020: 3-9.)

[26] Katsaggelos A K,Bahaadini S,Molina R. Audiovisual fusion: challenges and new approaches[J]. Proceedings of the IEEE,2015,103(9):1635-1653.

[27] 倪渊. 医疗知识图谱的构建及应用[R/OL]. [2021-03-04]. https://github.com/husthuke/awesome-

knowledge-graph/blob/master/conference. (Ni Y. Research on construction and application of medical knowledge graph [R/OL]. [2021-03-04]. https://github.com/husthuke/awesome-knowledge-graph/blob/master/conference.)

[28] Ellul C, Azzopardi J, Abela C. Notarypedia: a knowledge graph of historical notarial manuscripts[C]//OTM Confederated International Conferences "On the Move to Meaningful Internet Systems". Rhodes, Greece, 2019: 626-645.

[29] 雷洁, 赵瑞雪, 李思经,等. 知识图谱驱动的科研档案大数据管理系统构建研究[J]. 数字图书馆论坛, 2020(2):19-27. (Lei J, Zhao R X, Li S J, et al. Construction of knowledge graph for scientific research archives big data management system[J]. Digital Library Forum, 2020(2):19-27.)

[30] 申云凤,王英杰. 基于网络新闻语料的公共危机事件知识图谱构建[J]. 情报科学, 2021, 39(1):72-80. (Shen Y F, Wang Y J. Knowledge mapping of public crisis events based on Internet news corpus[J]. Information Science, 2021, 39(1):72-80.)

[31] 欧阳剑,梁珠芳,任树怀. 大规模中国历代存世典籍知识图谱构建研究[J]. 图书情报工作, 2021, 65(5): 126-135. (Ouyang J, Liang Z F, Ren S H. Research on the construction of knowledge graph of large-scale Chinese ancient books[J]. Library and Information Service, 2021, 65(5):126-135.)

[32] 冯惠玲. 学科探路时代——从未知中探索未来[J]. 信息资源管理学报, 2020, 10(3):4-10. (Feng H L. A path-finding era of discipline: exploring the future from the unknown[J]. Journal of Information Resources Management, 2020, 10(3):4-10.)

[33] 马费成,李志元. 新文科背景下我国图书情报学科的发展前景[J]. 中国图书馆学报, 2020, 46(6):4-15. (Ma F C, Li Z Y. Future prospect of library and information science in China in the context of new liberal arts [J]. Journal of Library Science in China, 2020, 46(6):4-15.)

[34] 孙建军,李阳,裴雷. "数智"赋能时代图情档变革之思考[J]. 图书情报知识, 2020(3):22-27. (Sun J J, Li Y, Pei L. Some thoughts on the reform of library, information and archives management in the era of data inteligence empowerment[J]. Documentation, Information & Knowledge, 2020(3):22-27.)

[35] 柯平. 新图情档——新文科建设中的图书情报与档案管理一级学科发展[J]. 情报资料工作, 2021, 42(1):15-20. (Ke P. New library and information archives: the development of the first class discipline of library information and archives management in the construction of new liberal arts[J]. Information and Documentation Services, 2021, 42(1):15-20.)

[36] 中华人民共和国数据安全法[EB/OL]. [2021-08-25]. http://www.npc.gov.cn/npc/c30834/202106/7c9af12f51334a73b56d7938f99a788a.shtml. (Data Security Law of the People's Republic of China[EB/OL]. [2021-08-25]. http://www.npc.gov.cn/npc/c30834/202106/7c9af12f51334a73b56d7938f99a788a.shtml.)

杨建梁 中国人民大学信息资源管理学院、数据工程与知识工程教育部重点实验室、中国人民大学电子文件管理研究中心师资博士后,讲师。北京 100872。

刘越男 中国人民大学信息资源管理学院、数据工程与知识工程教育部重点实验室、中国人民大学电子文件管理研究中心教授。北京 100872。

祁天娇 中国人民大学信息资源管理学院、数据工程与知识工程教育部重点实验室、中国人民大学电子文件管理研究中心师资博士后,讲师。北京 100872。

(收稿日期:2021-09-13)