

标准文献数字化转型的基础理论体系构建

夏磊 方思怡* 顾晓虹 蔡焱

(上海市质量和标准化研究院)

摘要: 标准数字化转型是标准领域的重要发展趋势。作为标准的重要载体形式,标准文献也面临数字化转型的发展需要。本研究从标准文献的自身特点与实际发展需要出发,提出了标准文献数字化转型的相关概念,基于此构建了标准文献数字化转型的基础理论体系模型,并围绕模型中的核心理论作简要阐述,为相关研究和工作提供一定的理论参考。

关键词: 标准文献数字化转型,标准文献,标准化,数据挖掘

DOI编码: 10.3969/j.issn.1674-5698.2024.01.006

Construction of the Basic Theoretical System for the Digital Transformation of Standards Literature

XIA Lei FANG Si-yi* GU Xiao-hong CAI Yan

(Shanghai Institute of Quality and Standardization)

Abstract: The digital transformation of standards is an important development trend. As an important carrier form of standards, standards literature also need to respond to the development needs of digital transformation. Starting from the characteristics and practical development needs of standards literature, this study proposes the concepts of the digital transformation of standards literature. Based on this, a basic theoretical system model for the digital transformation of standards literature is constructed, and a brief explanation is made around the core theories in the model, providing certain theoretical reference for related research and work.

Keywords: digital transformation of standards literature, standards literature, standardization, data mining

作者简介: 夏磊,硕士研究生,工程师,研究方向为标准信息管理。

方思怡,通信作者,硕士研究生,助理工程师,研究方向为标准数据挖掘、标准知识图谱。

顾晓虹,本科,高级工程师,研究方向为标准文献服务。

蔡焱,研究生,工程师,研究方向为标准文献服务。

0 引言

当前社会,数字化转型正前所未有地改变各行各业的生态面貌,标准领域也不例外。《国家标准化发展纲要》指出要加速推动标准化工作向数字化、网络化和智能化发展。近年来,标准数字化转型已逐渐成为标准领域的一大热点,旨在通过数字化手段来突破标准化工作的瓶颈进而实现数字化驱动标准化工作的转型升级^[1, 2]。

标准文献是指通过标准化活动,按规定程序经协商一致后制定的技术性文件,旨在为相关的活动提供规则和指南,具有可共同使用和重复使用的特性。作为标准化工作的重要载体,标准文献也同样面临数字化转型的发展需求。目前标准文献数字化转型在概念定义、范畴界定、基础理论体系、应用场景等方面仍存在大量的研究空白,其中基础理论体系是实现标准文献数字化转型落地应用的重要根基。本研究从标准文献数字化转型的实际需求和发展趋势出发,提出了标准文献数字化转型的相关概念,在此基础上构建了适用于标准文献的数字化转型基础理论体系模型,以期能为相关研究者提供一定的工作参考。

1 标准文献数字化转型的相关概念

1.1 标准文献数字化的定义

“数字化”(Digitization)是一种通过特定的信息技术将存储于非数字载体上的信息转化为计算机设备可识别和处理的数字信息的过程,其本质在于实现信息的数字化,数据是数字化的核心所在^[3]。

“标准文献数字化”是采用数字技术对标准文献内容和标准文献服务的全过程赋能,进而实现其在数字设备上读取、传输和使用的过程^[4]。与其他领域的数字化类似,标准文献数字化的核心也在于标准文献相关数据的数字化。

1.2 标准文献数字化转型的定义

“标准文献数字化转型”是指采用数字技术对标准文献的自身内容、服务模式和组织管理进行重构的变革性过程。作为一项复杂的系统性工程,

标准文献数字化转型涉及硬件设备投入、业务流程优化、组织管理、数字化人才培养等方面。

1.3 标准文献数字化转型的范畴

根据标准文献数字化转型的相关概念和实际需求,可以发现当前标准文献数字化转型的范畴主要包括标准文献自身内容的数字化、标准文献服务模式的数字化以及标准文献组织管理的数字化,其中标准文献自身内容的数字化是指通过数字技术将现有的纸质标准转化为机器可读取和操作的数字模态,标准文献服务模式的数字化是指标准文献核心业务的数字化重塑,主要包括标准文献编目加工、标准文献读者服务、标准文献数据挖掘等,标准文献组织管理的数字化则涉及标准文献数据、人员和管理方面的数字化。

2 标准文献数字化转型的基础理论体系模型

本研究从标准文献数字化转型的主要方面、主体对象和维度类型出发,将标准文献数字化转型的基础理论划分为“3方面-3主体-2维度”模型(简称“3-3-2”模型),在基础上形成适用于标准文献的数字化转型理论框架和技术体系。基础理论体系模型的概况如图1所示。

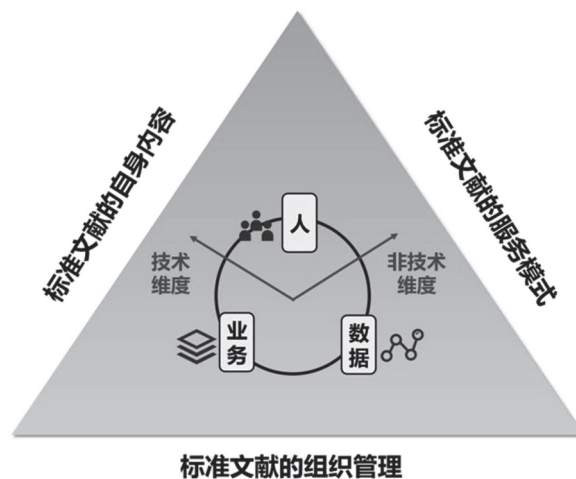


图1 标准文献数字化转型的基础理论体系模型

标准文献数字化转型“3-3-2”模型的“3方面”是指标准文献的自身内容、服务模式和组织管

理,“3主体”是指标准文献涉及的各类人员、数据和业务,“2维度”是指技术维度和非技术维度。

2.1 3个主要方面——“3方面”

2.1.1 主要方面I: 标准文献自身内容的数字化转型

标准文献的自身内容也即标准文献本身,标准文献内容的数字化转型是标准文献服务模式和组织管理数字化转型的重要基础,其目的旨在实现标准文献内容的数字化,通过一系列数字技术将纸质标准转化为机器可读取和操作的数字模态,由此突破标准文献的纸质形态^[5,6]。目前我国的标准文献大多以纸质文本的形式存储,从纸质文本到机器可读标准,需要历经文字识别、文本结构化、文本挖掘、知识组织和深度加工等多个流程。

2.1.2 主要方面II: 标准文献服务模式的数字化转型

标准文献服务主要包括标准文献编目加工、标准文献知识服务、标准文献产品研发等。标准文献服务模式的数字化转型也即在转变服务理念的基础上,通过数字技术重塑上述服务的流程和工具,实现服务模式的变革。

2.1.3 主要方面III: 标准文献组织管理的数字化转型

数字化转型的成败与组织的文化意识密切相关,因此管理变革是数字化转型的重要组成,通常需要遵循系统性、共享性、集成性和安全性的原则。本研究从标准文献数字化转型的管理需求和未来发展出发,将标准文献组织管理的数字化转型划分为战略管理、组织管理、数字化成熟度评估、数字化人才培养等方面。

2.2 3个主体对象——“3主体”

2.2.1 主体对象I: 数据

富有价值的数据是驱动数字化转型的关键动力^[7],也是布局数字战略的基础,因此“数据”是标准文献数字化转型的一大重要主体。标准文献数字化转型的相关数据主要有标准文献的文本数据、标准文献的业务数据、标准文献用户的行为数据等。

2.2.2 主体对象II: 人

“人”是标准文献数字化转型的制定者、参与者和执行者,故标准文献数字化转型的主体对象之一是“人”。根据角色差异,可以将标准文献数字化

转型的有关人员划分为标准文献用户、标准文献工作人员和标准文献管理人员,其中标准文献用户的需求是标准文献数字化转型的重要驱动力。

2.2.3 主体对象III: 业务

标准文献业务是标准文献的具体任务,业务流程的优化与升级是标准文献数字化转型的重要目标,因此本研究将业务列为标准文献数字化转型的主体对象之一。

2.3 两个维度类型——“2维度”

2.3.1 维度类型I: 技术维度

技术维度主要包括标准文献数字化转型所涉及的技术范式和其他技术应用相关的部分。

2.3.2 维度类型II: 非技术维度

非技术维度则主要是指标准文献数字化转型中与技术无关的内容。

3 标准文献数字化转型的关键核心理论

标准文献的不同业务板块相互联系,故基础理论体系模型中不同方面、主体和维度的理论并非完全分隔,而是存在一定程度的重叠。本研究梳理并总结了标准文献数字化转型中的关键核心理论,并做了简要说明,具体内容如下。

3.1 标准文献的光学字符识别理论

光学字符识别(Optical Character Recognition, OCR)是指将图像信息转化为计算机可识别字符的过程,主要包括界定文字字符所在区域、记录字符内容、确定字符识别的唯一性等流程^[8]。

鉴于当前国内仍处于纸质标准为主的局面,标准文献的OCR被视为构建机器可读标准和实现标准文献内容数字化的必由之路。标准文献OCR的实施流程可以被概括为预处理、文字区域检测与文字字符识别、优化处理这3个步骤^[9]。当前标准文献的OCR理论仍在发展阶段,存在基于工具和基于算法两大类OCR技术路线^[10],但两者在准确性上仍有较大的提升空间。

3.2 标准文献的结构化与数据规范理论

文本结构化是指将非结构化或半结构化文本转化为结构化文本的过程。标准文献是典型的非

结构化技术文本,标准文献内容的结构化处理是继OCR后又一个与标准文献内容数字化密切相关的流程。

自步入数字化转型时代以来,以数字对象描述为中心的文献信息资源描述框架开始成为大势所趋,主要包括基于资源描述框架(Resource Description Framework, RDF)模型的描述框架、基于实体关系(Entity-Relation, E-R)模型的描述方式等。作为图情领域的一大分支,标准文献的馆藏资源描述同样应当结合不同业务的需要建立以标准数字资源为中心的数据规范,形成适用于不同业务场景的数据框架,例如:在标准文献资源建设、编目和读者服务领域,可根据业务工作的实际需求制定相应的标准元数据体系,而对于有深度挖掘需求的标准文献应用研究领域,则需根据实际的业务目标,构建颗粒度更细、机器可读性更高的标准文献标注结构化数据集规范。

3.3 面向标准文献的自然语言处理理论

自然语言处理(Natural Language Processing, NLP)是指用计算机来处理、理解和运用人类语言的过程。标准文献是具有科技属性的自然语言文本,标准文献的加工与挖掘同样需要借助强大的NLP技术。

本研究深入分析了NLP通用任务在标准文献数字化场景中的应用前景,总结了NLP技术在标准文献数字化转型中的典型应用场景(详见表1)。

3.4 标准文献的数据分析与挖掘理论

数据是标准文献数字化转型的一大主体。标准文献数据是重要的基础性战略资源,标准数据分析与挖掘能为标准知识服务和组织管理提供必要的客观引导。

大数据分析的常用方法主要有关联分析(Association Analysis, AA)、分类分析、预测分析、聚类分析、可视化分析、离群点或孤立点分析等^[12, 13],其主要的流程包括数据预处理、数据挖掘和数据综合分析这3个步骤,分别对应数据层、知识层和应用层^[14]。从大数据的定义上来看,标准文献数据也属于大数据家族的一员。标准文献数据具有多源异构、数据量大、类型多样和价值密度大等特点。作为重要的经济资产和战略资源,标准文献大数据中蕴含许多有望带来经济效益的深度知识和价值^[15],梳理并形成适用于标准文献应用场景的数据分析与挖掘路径对于标准文献数字化转型的深入开展具有重要的实践性意义。

3.5 标准文献的知识组织理论

高效敏捷的标准文献知识组织能力是实现标准文献数字化转型的重要基础。标准文献的知识组织模式包括标准文献知识组织方式和存储工具两方面,前者主要是通过标准知识图谱来实现,后者则主要是指存储知识图谱所需的图数据库。

作为人工智能(Artificial Intelligence, AI)的重要组成,知识图谱(Knowledge Graph, KG)为数字化转型提供了从海量非结构化数据中抽取知识并通过图挖掘关联关系和实现知识组织的技术手

表1 NLP技术在标准文献数字化转型中的典型应用场景

NLP通用任务	具体说明	标准文献数字化转型中的应用场景	标准文献数字产品
机器翻译	将一种语言自动翻译为另一种或多种语言	标准文献智能翻译	标准文献多语言互译产品
智能问答	自动回答提问	标准智能问答	标准智能客服
文摘生成	自动归纳、总结并生成文本摘要	标准文献摘要	标准文献编目产品
文本挖掘	从非结构化文本中抽取结构化信息	标准命名实体识别、关系识别	标准指标识别与比对
文本分类	对采集而来的文本进行主题分析,进而实现文本的自动分类	标准文献分类	标准文献编目产品
文本聚类	在无需标记的条件下将相近距离的文本聚类在一类中	标准文献服务邮件过滤、用户问题识别	标准文献服务产品
舆情分析	判断当前舆论的导向	标准舆情分析	标准舆情分析产品
个性推荐	推荐与检索内容相关的信息	标准文献推荐	标准文献智能检索与推荐系统
情感分析	通过对主观性文本的情感极性进行识别来判断用户的情感倾向 ^[11]	标准文献用户分析	标准文献用户产品

段,被视为各行业产业链数字化转型的新引擎^[16]。近来的研究团队逐渐意识到了知识图谱技术在标准文献数字化转型中的重要性,陆续形成了标准文献知识图谱的初步成果,但截至目前,尚未形成业内统一的成熟理论体系。

3.6 标准文献的数据治理理论

数据治理(Data Governance, DG)是指在确保数据安全的条件下,通过一套特定的规则对数据的全生命周期进行全面而有序的规范化监管以优化数据配置、实现数据合理流动并释放其最大潜在价值的过程^[17, 18]。

数据治理是标准文献数字化转型中不可或缺的一环。当前标准文献的数据治理普遍存在数据治理体系缺乏、战略意识欠缺、数据孤岛、数据质量不高和数据安全风险隐患等问题^[19],亟需构建适用于标准文献的数据治理体系和数据标准。

3.7 标准文献的产品设计与用户行为分析理论

标准文献服务模式的数字化转型是标准文献数字化转型的重要组成。从广义上来看,每一种标准文献服务都可以被视为一种标准文献“产品”。标准文献服务模式数字化转型的最终目的在于为标准文献用户提供智慧化的定制服务,解决以往标准文献用户获取标准文献信息的局限性,包括获取高质量的标准文献知识、帮助用户在海量数据中快速定位个人所需内容、根据用户实际需求定制个性化知识等^[20],在这其中,产品分析(Product Analysis, PA)是实现标准文献服务模式智慧化的重要基础。在标准文献的产品分析框架中,标准文献用户分析是一大重点,构建适用于标准文献的用

户分析和用户画像理论对标准文献服务模式的数字化转型至关重要。

3.8 标准文献数字化转型的成熟度评估理论

正确评估数字化转型的成熟度对标准文献数字化转型的顺利推进至关重要。构建科学合理的标准文献数字化转型的成熟度评估模型能够便于标准文献的相关从业者判断标准文献所处的数字化阶段,找出不利因素,及时查漏补缺,制定相应的对策。已有的研究大多围绕标准数字化转型开展,而标准文献数字化转型的成熟度评估仍存在大量空白。

3.9 标准文献数字人才的培养与发展理论

人才已成为数字经济的核心生产要素,标准文献数字化转型的核心是人的思维方式、成长模式以及工作模式的转变^[21]。当前深入推进标准文献数字化转型的一大难题在于标准文献数字人才的短缺^[22],如何整合现有人力资源、培育一流的标准文献数字人才已成为标准文献数字化转型顺利推进的重要课题。

4 结语

本研究根据标准文献的现状和发展趋势提出了标准文献数字化转型的相关概念,基于此构建了标准文献数字化转型的基础理论体系,形成了标准文献数字化转型基础理论体系的“3-3-2”模型,并重点罗列了其中的核心基础理论,在后续工作中将围绕此模型开展标准文献数字化转型的相关研究与应用,进一步提高标准文献的数字化发展水平。

参考文献

- [1] 王立玺,吕千千,牛艳茹. 标准数字化工作关键路径探究[J]. 信息技术与标准化, 2022(10):27-30+42.
- [2] 崔静,杨建军. 标准数字化服务探索[J]. 信息技术与标准化, 2022(10):6-12.
- [3] 刘耀宏. 传统企业数字化转型实践[J]. 信息技术与标准化, 2022(06):24-28.
- [4] 张程,龚庆,姚波,等. 机器可读标准实现方法研究[J]. 中国标准化, 2022(21):56-61.
- [5] 卢铁林,汪烁,樊子天,等. 标准数字化等级评价视角的转型模式研究[J]. 信息技术与标准化, 2022(10):31-35.
- [6] 马超,邓桃,周勤勇,等. 面向电力领域的标准数字化转型工作研究——需求分析、转型路径与应用场景[J]. 中国标准化, 2022(23):87-92.
- [7] 赵发珍. 图书馆治理数字化转型:理论内涵、内在机理及推进路径[J]. 图书馆学研究, 2023(03):2-8. DOI:10.15941/j.cnki.issn1001-0424.2023.03.001.

- [8] 史素霞,常婉秋,宋志英. 基于UCI数据集的OCR光学字符识别[J]. 科技创新与应用,2022,12(35):50-53. DOI:10.19981/j.CN23-1581/G3.2022.35.012.
- [9] 陈心怡,张华,方伟,等. 光学字符识别与标准数字化转型[J]. 中国标准化, 2022(20):56-59.
- [10] 方思怡,夏磊. NLP技术在标准文献挖掘中的应用与问题初探[J]. 中国标准化, 2022(17):62-66.
- [11] 熊回香,杨梦婷,李玉媛. 基于深度学习的信息组织与检索研究综述[J]. 情报科学,2020,38(03):3-10. DOI:10.13833/j.issn.1007-7634.2020.03.001.
- [12] 陈欣,曾红月. 数据挖掘的分析方法综述[J]. 福建电脑, 2004 (01):31-32.
- [13] 任磊,杜一,马帅,等. 大数据可视分析综述[J].软件学报, 2014,25(09):1909-1936.DOI:10.13328/j.cnki.jos.004645.
- [14] 苏怀,张劲军. 天然气管网大数据分析方法及发展建议[J]. 油气标准储运, 2020,39(10):1081-1095.
- [15] 王万良,张兆娟,高楠,等.基于人工智能技术的大数据分析方法研究进展[J]. 计算机集成制造系统, 2019, 25(03): 529-547.DOI:10.13196/j.cims.2019.03.001.
- [16] 本刊编辑部.知识图谱: 打造数字化转型新引擎 推动全产业链落地应用[J].信息技术与标准化,2021(05):1.
- [17] 林伟,周耀铭. 国内外数据治理研究述评[J]. 数字图书馆论坛, 2022(06):65-72.
- [18] 黄国平. 创新和重塑数据治理体系——以金融数据治理为例[J]. 经济管理, 2023,45(01):25-42.DOI:10.19616/j.cnki.bmj.2023.01.002.
- [19] 崔永梅,张莉莉,赵晓丽,等. 构建数据资产组织赋能体系推动企业数字化转型[C]//中国企业改革与发展研究会.中国企业改革发展优秀成果2020(第四届)下卷. 北京: 中国商业出版社(China Commerce and Trade Press),2020:460-469.DOI:10.26914/c.cnkihy.2020.061849.
- [20] 史新蕾. 例谈期刊编辑的产品经理思维能力培养[J]. 科技与出版,2023(04):91-95.DOI:10.16510/j.cnki.kjych.20230419.003.
- [21] 吴江. 面向数字转型的人才发展新定义[J]. 中国科技人才, 2021(04):3.
- [22] 张一博. 数字经济背景下长春市人才发展对策研究——基于上海、杭州数字经济人才发展的分析[J]. 中国市场,2022(06):174-175.DOI:10.13939/j.cnki.zgsc.2022.06.174.