数据结构化技术与目录学方法论的本质辨析*

彭贤哲　郑建明　石　进

【摘　要】明晰数据结构化技术与目录学的本质联系，有利于为技术的发展提供坚实的理论基础，丰富与完善现有目录学的方法理论体系。文章通过回顾目录学方法在当下技术影响下的传承与变化，透析数据结构化技术中的目录学方法原理，展现二者互为表里、相互促进的共生耦合关系，进而厘清二者的潜在联系。自动化技术手段的引入推动了目录学方法论的现代化，目录学方法原理的借鉴促进数据结构化技术的诞生与发展。文章以一种整合的视角辨析二者关系，认为数据结构化技术是现代信息技术对目录学方法论的一种借鉴、继承、发扬与完善。

【关键词】目录学方法　数据结构化技术　文献揭示　文献组织

Abstract: A clear understanding of the essential relationship between data structuring technology and bibliography is conducive to providing a solid theoretical foundation for the development of technology, and enriching the existing methodological and theoretical system of bibliography. In order to clarify the potential relationship between the two, this study reviews the inheritance and change of bibliography method under the influence of current technology, analyzes the principle of bibliography method in data structuring technology, and shows their symbiotic coupling relationship. The introduction of automatic technology promotes the modernization of bibliography methodology, and the reference of bibliography principle promotes the innovation and development of data structuring technology. This research analyzes the relationship between the two from an integrated perspective, and holds that data structuring technology is a kind of reference, inheritance, development and improvement of modern information technology to bibliography methodology.

Key words: bibliography methodology　data structuring technology　disclosure of literature　organization for literature
DOI:10.15941/j.cnki.issn1001-0424.2024.02.012

0　引言

随着数字时代的到来，目录实践已逐渐加入到计算机应用、数据挖掘、数据处理等现代信息技术中，如检索语言、控制词表、主题词表、元数据、关联数据等目录学工具已被广泛推广与使用，这使得目录学的致用属性在新时代得到了重塑与加强[1]。虽然这些技术或多或少应用了目录学的方法工具，但在技术的发展演化过程中由于迭代更新、优化完善，技术的命名或展开过程中逐渐隐匿了其中蕴涵的目录学原理及方法论，或修改名称以体现工具改善过程，或省略基础原理介绍以展现技术创新点。在技术的这种命名习惯促使下，目录学很难直接体现在新兴信息技术中，但又确实“暗藏”于其底层的技术原理之中，其社会应用价值未能得到直接有效的彰显[2-3]。

长久以来，关于信息技术体系的研究多从计算机科学这一独立视角展开，注重优化更新已有技术以提高效率、优化服务，包括各类算法模型的迭代更新、数据处理工具的改进完善等，如用于节省计算成本和提高计算效率的云计算技术[4]、改善交互体验和机器理解能力的大语言模型[5]、增强数据安全性和开放性的区块链技术[6]等，基本关注于提升技术实施过程中某一具体环节的效率或性能，对这些技术的底层原理尚未展开深入的探讨。近年来，已有部分研究结合目录学与计算机科学两种视角，指出目录学思想作为大数据管理的理论基础，在数据结构化的宏观过程中得到了传承和应用，以个别信息技术为例，初步探索了信息技术的底层原理。如：石进等[7]提出目录

* 本文系国家社会科学基金项目“面向国家安全的科技情报态势感知研究”（项目编号：21BTQ012）的研究成果之一。

学的理论方法可用于解决当下海量数据资源的揭示与报导问题，并认为这是目录学在大数据时代的新使命；熊翔宇和郑建明[8]以图书馆数字信息资源重构过程为实例，说明了大数据管理在于完成目录工作的实质——知识资源的组织与整序，论证了大数据管理过程中蕴藏的目录学思想；彭贤哲等[9]将数据管理过程中数据结构化步骤作为研究对象，从宏观上辨析了该步骤的本质过程、书目机理和书目思想，印证了数据结构化过程是对目录学思想的传承与应用。技术化和知识化作为现代目录学发展的两大特征[10]，以信息采集技术、信息组织技术、信息检索技术、信息传播技术、信息共享技术等组成的信息技术体系[11]，对目录学的方法理论广泛应用，为海量、异构、异源的信息资源整理、规范、导航、发现和管理提供了现实可能性[12]。

如今，目录学知识虽然广泛应用于各个领域，但却隐匿于各种工具技术的底层原理之中，现实中已并非是实践落后于理论的问题，而是理论远远落后于数字时代下的实践工作。因此，为解决理论与实践失衡与脱节的问题，有必要弘扬古典目录学书目实践与学术研究相统一的优良传统[3]，从实践工作中挖掘、提炼、总结、应用蕴藏于具体任务场景中的目录学理论。为此，本研究拟以当前广泛使用的数据结构化技术为研究对象，从具体细节层面剖析目录学方法论在技术上的发展与应用，更为直观地展现数据结构化技术对目录学思想方法的传承与发扬，丰富与完善现有目录学的方法理论体系，由表及里地挖掘、印证、梳理数据结构化技术中蕴含的原理与目录学的本质联系。

1 数据结构化技术

认识数据结构化技术的核心本质，需要围绕它的应有含义、组成要素、表现形式 3 个方面逐层展开。从这 3 个方面解析数据结构化技术的实现目标、具体任务、实施过程、工具方法，有助于归纳整理出它的核心本质。

1.1 数据结构化技术的含义

根据数据结构的类型不同，可将数据分为结构化数据、半结构化数据和非结构化数据。其中，将无固定结构的半结构化、非结构化数据转化为具备特定结构数据的有序化过程，称为数据结构化。数据结构化技术是指在数据结构化过程中应用的具体方法和手段，包括但不限于创建索引、信息抽取、自动分类、构建主题词表等。在将文本、图像、音频等非结构化数据转化为便于管理利用的结构化数据的过程中，该技术旨在提高数据结构化过程的自动化程度和效率，影响最终转化获取的结构化数据的准确性和质量。

1.2 数据结构化技术的组成要素

非结构化数据作为现有数据类型的主体，具有处理难度大、信息丰富、使用价值高的特点。现有大量关于数据结构化技术的应用探究，为这类数据的利用管理和价值挖掘提供了行之有效的手段途径，并积累了各种形式不一的方法。

应用在数据结构化过程的数据结构化技术，可大致由 5 项特定的结构化处理任务（表 1）所组成，分别为抽取识别非结构化数据中的分割单元、整合关联识别抽取的分割单元、根据分割单元特征实现分类或聚类、依据特征属性或关联关系融合、重组不同分割单元。其中，与目录学中主题法是以文献中论述对象为标注组织单元类似，非结构化数据中处理单元的识别抽取可用人工手段、机器学习或深度学习实现，并进一步标注分割单元的特征信息，形成针对非结构化数据的二次描述数据；基于产生的描述数据，分割单元的各项描述特征间的共现关系、原始数据资源的引用链接关系、语义内容特征的相近或相似关系可作为分割单元建立关联关系的基础。之后，建立关联关系的分割单元在分类算法、聚类算法基础上，可进一步依据描述的特征信息实现自动化聚集；聚集在相近位置的同类分割单元一般难免存在一定的信息冗余，计算分割单元描述特征或描述关系之间的相似性矩阵，可实现单元特征或关联关系的融合；最终，抽取识别的分割单元根据分类法、主题法等组织方法完成序化重组。

表 1 数据结构化技术的组成要素

组成要素	研究方法或角度	具体过程或方法	代表性研究
数据单元抽取	人工手段	以方法、工具、资源、理论等知识实体为分割单元，通过内容分析法、人工标注方式实现	[13-14]
	机器学习	支持向量机（SVM）、条件随机场（CRF）等	[15-16]
	深度学习	卷积神经网络（CNN）、循环神经网络（RNN）、长短期记忆模型（LSTM）以及注意力机制等	[17-18]

组成要素	研究方法或角度	具体过程或方法	代表性研究
单元关联整合	特征共现	分割单元间的特征属性的重合、共存、耦合等关系，构建共现网络，实现其在多个维度的关联聚集	[19]
	引用链接	资源间的相互引证或指向关系，依据此类关系建立的网络，可实现信息资源单元在不同传承演变支脉上的有序聚集	[20]
	语义关联	注重从内容层面关联聚集信息资源，并可进一步通过知识推理等方法发现潜在关联关系	[21]
单元分类聚类	抽取单元分类	向量机、朴素贝叶斯、决策树和随机森林	[22-23]
	抽取单元聚类	基于划分、基于层次、基于密度、基于网格的聚类方法，此外还包括一些新兴方法，如粗糙集方法和模糊聚类方法	[24]
抽取单元融合	特征属性	字段映射、字段拆分、数据记录滤重、异构数据加权、元数据可扩展	[25-26]
	关联关系	将不同单元关系有效融合成一个新关系数据来表征单元之间的关联特征，以多源信息的相似矩阵或距离矩阵表示	[27]
抽取单元重组	信息组织	字顺法、分类法、主题法等	[28-29]

1.3 数据结构化技术的表现形式

数据结构化过程中蕴涵有关联、索引、标引、组织 4 种书目机制[3]，依托计算机处理与服务工具，数据结构化技术展现为数据关联、数据索引、数据标引、数据组织 4 种表现形式（表 2）。

表 2 数据结构化技术的表现形式

表现形式	典型技术	具体过程及功能作用	代表性研究
数据关联	元数据	关于数据的数据或描述数据的数据，是提供关于信息资源或数据的一种结构化数据，是对信息资源结构化的描述数据，已应用在数据库管理系统（DBMS）和网页构建过程中	[30]
	语义网	通过资源的语义描述，计算资源语义描述数据之间的语义相似度，连接离散的、不同类型、不同结构的资源，形成紧密的、结构化的资源关联网络	[31-32]
数据索引	稀疏索引 稠密索引 正排索引 倒排索引	选取有检索意义的主题信息作为标目，并在其后附上位置信息；依据拼音或笔画等规则排列、合并标目，并把其后的位置信息按出现顺序依次接续。应用潜在语义分析、自动分词等技术，辨析同义词和异义词，提高资源查询过程的检准率和检全率	[33-34]
数据标引	抽词标引	直接从原始资源中抽取词或短语作为标引词，用以描述资源主题内容的过程	[35]
	赋词标引	根据资源的内容特征，从受控词表中选择叙词或主题词作为资源检索标识的过程	[36-38]
数据组织	知识图谱	以结构化形式表示现实世界中存在的实体及实体之间的语义关系的语义知识库，旨在描述资源中知识实体、知识概念及其关系，通过推理机制深入挖掘资源中存在的隐性知识，实现知识发现与知识服务	[39-40]
	主题图	由主题、关联和资源出处 3 个要素构成，描述出资源之间的关系，可定位于知识概念关联的资源位置，也可描述知识概念之间的语义联系，关联主题概念和资源实体，实现资源在位置和内容上的聚集	[41-43]

其中，数据关联技术以建立数据集合之间的关系或联结为途径，将离散的、无规律的数据资源按特定的规则形成关联体系，便于资源的系统组织、管理与利用，包括元数据、语义网等典型技术；数据索引技术以构建揭示数据内容的位置信息为目的，可提高非结构化数据的查检效率，形成诸如稠密索引、稀疏索引等多种索引形式；数据标引技术在于提取原始非结构化数据中蕴含的高价值特征信息，以提取的结构化特征数据揭示、描述并替代非结构化数据，可分为抽词标引和赋词标引两种类型[35]，是管理利用非结构化数据的关键所在；数据组织技术通过按一定的方式和规则对数据进行归并、存储和处理，最终形成清晰严密的数据管理系统，例如知识图谱、主题图技术等。

1.4 数据结构化技术的核心本质

虽然数据结构化技术具体实施过程的表现形式各不相同，但其本质过程是一致的，可大致将其分为抽取识别非结构化数据中的分割单元、描述分割单元的特征实现单元的关联整合、重塑非结构化数据组织结构完成序化存储 3 个步骤[3]。这个本质过程展现在具体的数据结构化技术的方法过程中，或呈现为对这 3 个过程的部分组合，或表现为对其中某个过程的具体细化。结合数据结构化技术的 5 个组成要素和 4 种表现形式，数据结构化的本质过程具体落实在数据结构化技术中时，其具体执行过程亦可分为 3 个步骤：（1）识别非结构化数据分割单元，抽取其各项特征信息揭示原始数据；（2）通过二次数据的表示形式实现分割单元的特征描述、相似度计算、类别判定等；（3）以二次描述数据为计算依据，完成分割单元的关联、融合、重组、序化。

2 目录学方法论

在不同国家、不同历史时期的目录工作实践过程中，前人已提出、整理、归纳了一系列行之有效的目录学方法。在此基础上，有关目录学方法论核心本质的梳理总结工作，可以从方法论的含义、构成要素、具体示例、执行过程等多个方面逐步开展。

2.1 目录学方法论的含义

方法论作为读书治学的工具，并非简单地指某个学科的个别方法的特殊应用，而是一种关于思维方法的学说，通过对不断增长的科学知识和程序的反思而实现知识领域的扩大，是一个反映该科学知识的系统，允许建立一套方法论规范，以满足科学工作者对某一类方法的要求[44]。方法论作为一套完整的实践引导体系，是指导实践的理论基础，也是从社会实践中总结和摸索出来、符合客观发展规律、符合人们对研究对象进行科学分析的实践步骤[45]。

目录学方法论作为目录学理论与书目工作实践的中介，是目录学理论与实践产生联系且相互作用的核心纽带。目录学方法论来自目录实践工作，指实际应用于文献资源的整合序化过程中形成的一系列方法集合，为后续书目工作实践的顺利开展提供先例参考和借鉴对象，为目录学理论体系的应用与发展提供具体可行的执行方案和落实途径。在新的时代背景下，有必要重新探讨、分析与发展目录学方法论体系，促进人文和技术的新融合，为目录学研究开拓更广的领域，创设更大的发展空间[46]。

2.2 目录学方法论的构成要素

方法论通过对不同学科方法的应用、完善、概括、反思与总结，虽然不能脱离各具体学科而独立存在，但可以从各具体学科的方法论体系中汲取营养，丰富自己，在普遍性上高于各具体学科。根据方法论的定义，目录学方法论属于具体科学方法论，其一表现为目录学方法及其体系建立的学说，其二表现为目录学诸方法的总和，可视为目录学所采用的方式方法相互关联而构成的一个统一整体[47]。目录学方法论作为方法论在特定学科的进一步细化，应符合方法论形成的条件，即至少包含思想理论、概念模型、基本原则和规则、过程和程序、具体方法、评估标准 6 个构成要素[48]。

细论目录学方法论的组成要素见图 1。目录学思想是目录学方法论的理论基础，包括客观著录、书目编纂、文献分类等思想，是指导各种方法、技术和工具等应用的理论依据；概念模型在于揭示目录学方法论的研究问题及解决方案，如文献描述、文献关联、文献分类等模型，阐明目录实践工作中的研究问题及研究逻辑；基本原则和规则主要指各项文献整理工作应遵循的事项，明确在不同情境下的必要事项和禁忌事项，例如，著录规则限定了著录的必要项和非必要项等，保证过程的合理性和目标的符合性；根据概念模型展开的具体过程和程序，则是为解决研究问题展开的具体步骤；具体方法是目录学方法论的核心，用于指导目录工作的具体过程，决定目录工作对预期目标的完成度和质量水平；目录学方法论中的评估标准主要指目录工作对文献揭示、整合、序化等各项目标的实现程度，据此产生反馈，用于完善丰富预设的概念模型、优化更新已有的具体方法等。

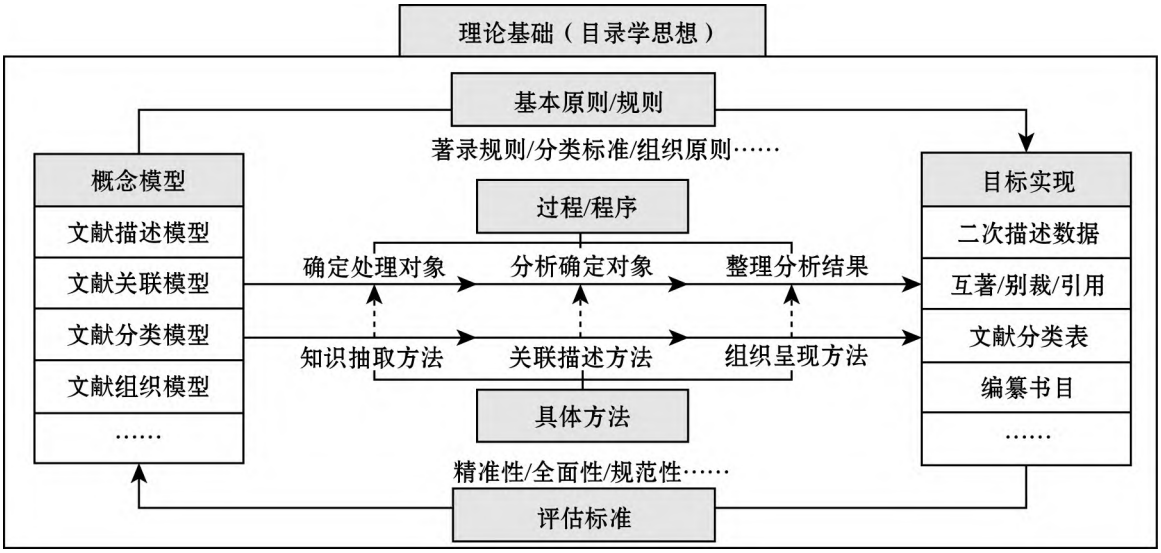

图 1　目录学方法论的组成元素

2.3　目录学方法论的核心及本质

目录工作的示例方法、具体过程及本质见表 3。

表 3　目录工作的示例方法、具体过程及本质

目录学方法	具体示例方法	具体过程	本质
文献揭示方法	著录	按照一定的著录规则制定相应的描述项目，并规定项目的排列顺序、标识符号形成二次文献，完成对原始文献的内容特征和外表特征进行客观描述的过程	揭示文献内容和形式特征
	提要、序跋	解释文献题意、介绍作者生平与学术思想以及揭示文献内容和评价学术得失	揭示文献内容和影响特征
	综述	分析、综合、提炼原始文献，编写出全面反映某学科或领域及专题的发展现状、存在问题及发展趋势	揭示文献之间的关联特征
	书评	分析著作的优劣得失，评议其思想性、科学性、知识性、实用性和影响力	揭示文献内容和影响特征
文献组织方法	索引	将特定范围内的文献资料中特定主题或线索摘录出来并注明位置信息，按一定次序编排便于用户查检	揭示文献内容和位置特征
	分类法	将学科性质相同或相近的文献排在相近的位置，可系统地揭示文献的本质属性和内容上的关联关系，由此反映该学科的文献分布状况	基于文献整体性的内容特征实现零散文献的关联聚集组织
	主题法	把论述同一主题（或事物对象）的文献加以集中，可适应用户对文献资料更为具体、更为个性化的查检利用需求	基于文献细节性的内容特征实现零散文献的关联聚集组织
	书目控制	通过编制和利用索引、摘要等二次文献，用以了解某一主题领域、特定类型、某个门类的文献集群，并掌握其分布和发展规律	基于文献形式或部分内容特征实现零散文献的关联聚集组织

根据目录学方法论的组成要素分析，深入了解目录学方法论，需要认识分析其中的核心组成部分，即实践工

作过程中形成的具体方法，包括中国目录学和西方目录学在不同历史时期提炼、继承、发扬的各种工具技术等。中国目录学和西方目录学在文献整理过程中，二者由于目的不一致，因而在研究方法的选择上各有侧重。中国古典目录学强调依据文献内容特征实现分类，达到“辨”与“考”的目的，重视利用小序法和提要法揭示文献的语义与语用特征。西方目录学着眼于“方便地获取图书”，注重从形式特征和部分内容特征上编撰二次书目数据，形成索引并排序，倚重书目控制的方法快速找寻到与检索信息字面匹配的图书[49]。总结二者共有的本原，其工作方法仅是实施策略和角度有所不同，但均是在描述文献整体或局部上的内容、形式、位置、关联等各项特征，形成二次文献用于描述原始文献，并以此为依据实现原始孤立文献资源的重组融合、关联聚集、查检利用。

在漫长的历史发展过程中，目录学积累了一系列行之有效的方法，包括校雠目录之法、书序解题之法、辑录注释之法、互著别裁之法、类例类序之法等[50]。当下此类传统工具方法与新兴信息技术的结合，相继产生了计算机编目、文献检索、书目情报等新思路、新方法，丰富了目录学科本身原有的方法论体系。技术的引入虽然催生或改变了目录学的一系列工作方法和呈现方式，但并未改变目录学方法的本质过程，现有的目录学方法论体系仍可分为文献揭示方法和文献组织方法两个大类（图 2），前者的本质在于以原始文献的各项特征为中心，形成揭示原始文献的二次文献，后者则聚焦于二次文献中提及特征之间的关联关系，据此序化整合原始文献资源。

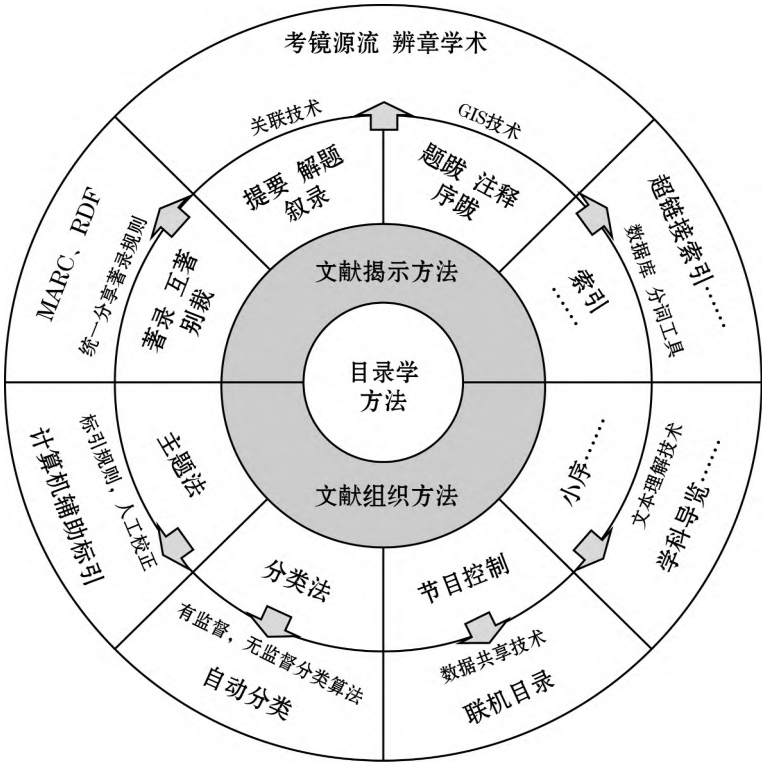

图 2 传统目录学方法在技术手段下的演变

对比中外，纵观古今，目录学方法论的本质基本可归纳为 3 点：（1）从整体或局部角度揭示原始文献的形式、内容、影响、关联等各项特征；（2）以二次文献形式精炼有效地描述或表示原始文献信息或文献之间的关联信息；（3）根据二次文献中提及的特征实现原始文献的关联、融合、重组、序化。

3 目录学方法论与数据结构化技术的本质联系

数据结构化处理作为大数据管理过程中的关键环节，其中蕴涵的目录学思想、目录学机理、书目机制在宏观层面已被澄清阐明[9]，但二者在实现数据或文献有序化过程的方法策略是否有所关联，还有待进一步辨析与论证。为此，剖析目录工作形成的方法论本质，对比数据结构化技术的底层原理，可展现目录学方法论和数据结构化技术在实施过程中的相通与相似之处，辨析二者潜在的本质联系。

3.1 研究对象的一致性

数据结构化技术的研究对象主要包括文本、图像、音频、视频等多模态异构数据，目录工作主要以各类文献为研究对象，包括期刊论文、专著、学位论文等，二者虽然外在表现形式不同，但具有一致的本质特点。首先，

二者的载体形态具有多源异构性的特点，文献的载体形态包括但不限于甲骨、青铜、缣帛、简牍、纸张、金石等，非结构化数据包括的数据范围更为多样，包括各类存储形式的多媒体数据等等；其次，二者的表现形式具有混合杂乱的特点，表现为格式多样且不统一，无固定标准，结构化程度较低，有待于数据结构化技术或目录工作的描述著录；最后，二者的关联方式具有单一浅显的特点，如原始文献之间仅存在引用等浅层次的位置关联关系，非结构化数据之间的关联方式也多以简单的链接引用形式出现，对于内容上的语义关联较为欠缺且不显著。

数据结构化技术和目录工作虽然在具体的研究对象上存在区别，但二者的研究对象在载体形态、表现形式、关联方式上均具有共性特点，这为二者在研究方法、执行步骤、预期目标等多个方面互通互鉴提供了先决条件。

3.2 研究方法的互通性

数据结构化技术与目录学方法论在研究方法方面是否具有本质关联，需要解析各类具体的数据结构化技术的原理和实施过程，从本源上论证二者在研究方法的互通性。如图3所示，由关联、索引、标引、组织4种书目机制衍生的4种数据结构化技术的具体实施过程，大部分体现或继承了目录学方法论包含的具体方法，其中以数据索引技术最为直接，在技术名称上则展现了索引构建方法在数据结构化技术中的应用与拓展，其他技术则需进一步剖析展开方可展现。

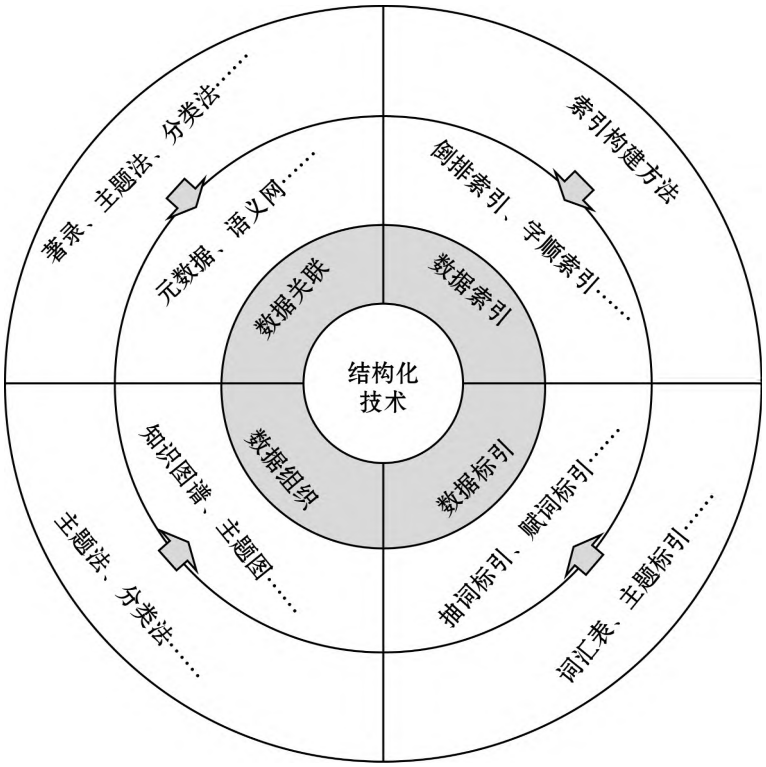

图3 数据结构化技术的目录学方法原理

数据关联技术通过建立数据集合之间的关系或联结的手段，将离散的、无规律的数据资源按特定的规则形成关联体系，便于对资源的系统组织、管理与利用。在此过程中，诞生了如元数据、语义网等典型的数据关联技术。以元数据为例，虽然其形式多样，但其仍是一种描述揭示数据资源的方法，与目录学的著录方法同样是一种针对描述对象的揭示方法，二者均是从描述对象的各项特征展开，为满足用户特定需求产生了由一系列特征组成的描述项集合，区别仅在于因用户需求不同而产生的描述项差异。

古典目录学的研究客体与起步工作是对文献的描述与标引，具体的标引工作主要出现在利用主题法组织文献的过程中，通过将文献中论述的各个事物对象用规范化的术语标引出来，作为后续组织、检索文献的标识，这与数据标引技术的具体工作是相通的，仅是实现手段由人工操作转化为计算机处理，在效率上获得了提升，但标引质量却仍有待于进一步提高。从目录学角度审视数据标引技术，其仅是标引主体和标引客体发生了变化，标引主体由主观性强、智能化程度更高的标注人员转向客观性强、处理速度更快的计算机，标引客体由传统的文献资源转变为更为泛化的数据对象，标引的本质过程仍是别无二致的。

相比传统的文献组织方法，数据组织技术拥有更为多样的组织形式、更为灵活的组织过程、更为高效的组织工具、更为实用的组织结果，但组织原理仍与目录学文献组织的原理是一脉相传的，即是在揭示描述对象特征的基础上，或采取分类法聚集相似特征的对象，或采用主题法关联论述同一实体的对象，或组合使用诸如索引等多种目录学方法分析对象特征完成组织过程，仅是实现过程由人工整理转向计算机处理，无法掩盖其是目录学传统文献组织方法的“现代化应用”的本质。

数据结构化技术利用现代化的各种工具手段，通过多样的揭示、描述、关联、组织过程，实现了非结构化数据的结构化分析、存储、管理、利用，虽然处理对象广泛，表现形式多样，呈现结果多变，但拨开纷繁复杂的各种结构化技术的执行过程来看，其采取的研究方法与目录学方法论在本质上具备互通性的特点，并对目录学方法论进行了一定的继承、修缮与发展，展现了目录学在当下环境的致用性特点。

3.3 本质过程的相似性

目录学作为一门历史悠久的学科，具备导读、揭示和组织文献信息等功能，自产生以来在人类文化传承过程中发挥了巨大的作用[51]。在信息愈发多样复杂的环境下，各种数据结构化技术在诞生之初就包含了目录学的工作方法原理，但在后续的迭代更新过程中，目录学的属性特点多被隐匿，无法直观地从技术的名称和简介中看出目录学的“身影”，但剖析各项数据结构化技术的本质过程，其仍是对目录学方法的一种传承和发展。

目录工作以原始文献资源作为处理对象，通过著录、提要等文献揭示的目录学方法，提取其中蕴含的形式、内容、位置、影响、关联等各项特征，并进一步过滤筛选形成描述原始文献的二次文献，如综述、索引、文摘、著录等，继而根据二次文献采取分类法、主题法等文献组织方法融合、关联、重组、序化原始文献资源，形成便于管理利用文献资源的结构化系统。数据结构化技术主要面向非结构化数据的处理加工，结合用户需求通过实体识别技术确定非结构化数据中的基本分割单元，并利用自动标注技术或自定义提取规则的方法获取分割单元的各项特征，形成描述分割单元的二次数据，如数据特征的矢量表达式、元数据等，结合二次数据通过与目录学相似的组织方法实现分割单元的融合、关联、重组与序化，最终获得表示非结构化数据资源的结构化数据系统。

如图 4 所示，目录学方法与数据结构化技术虽然在处理对象、具体过程、最终结果 3 个方面存在一定的差异，但二者在方法原理层面仍存在很多相似之处。首先，二者选用的揭示方法在揭示客体对象上存在差别，分别为原始文献和非结构化数据中的分割单元，但揭示方法的落脚点均在于提取揭示客体对象的各类属性特征；其次，二者均采用二次文献或二次数据的形式展现待组织对象的特征，并将其作为后续组织过程中的主要依据；最后，二者具备大致相同的组织方法，组织过程不外乎融合、关联、重组、排序 4 种形式，仅研究角度会存在区别，如文献资源之间与数据分割单元之间的关联维度不一致等。

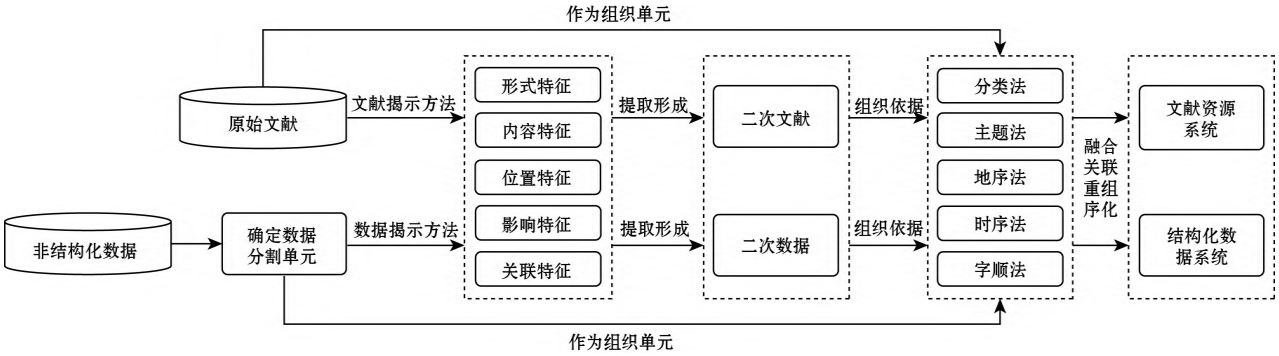

图 4 数据结构化技术与目录工作方法的本质过程

数据结构化作为目录学方法在计算机领域的另一种表现形式，由此衍生的技术在本质原理上与目录学的方法体系是互通互联的，二者的本质过程是基本相似的，且在底层原理上密切联系，可相互借鉴发展、优势互补，从而有助于数据结构化技术的优化更新、目录学方法论的完善发扬。

3.4 目标导向的趋同性

数据结构化技术和目录学方法论在研究对象上具有相似的特点，且具备互通相连的研究方法，详细展开的程序或过程在本质上也是一致的，这得益于二者在多个环节的目标需求是显著趋同的。

针对多源异构的文献资源或非结构化数据，二者均需明确待分析处理的研究对象，抽取识别其中隐含的处理单元，如文献记录过程中确定记录单元、非结构化数据处理过程中抽取目标对象等，目标均在于从多源异构的载

体中抽取符合用户需求的相同结构的分割单元，便于后续的分析、管理与应用等过程。再者，利用文献描述方法、数据矢量表达公式等研究方法揭示分割单元的各项特征，达到以结构化的二次数据或二次文献揭示描述格式多样、无序杂乱的原始数据或原始文献的目的。此外，非结构化数据或原始文献资源之间的关联程度较浅，依据二次描述文献或数据，还需通过分类法、主题法等关联方法加深原始资源之间的关联程度；在揭示原始资源、深化原始资源之间的关联程度的基础上，数据结构化技术和目录学方法论的最终目的在于序化组织原始资源。

剖析数据结构化技术和目录学方法论的详细步骤，二者在各个环节均具有相近的目标导向，包括以标准结构化的二次资源形式揭示原始资源的需求、通过二次资源的各项描述特征深度关联原始资源的需求、根据关联描述信息序化组织原始资源的需求等。数据结构化技术和目录学方法论在目标导向上的趋同性，是二者在研究方法和执行过程产生本质关联的基础。

3.5 发展历程的关联性

如图5所示，数据结构化技术的出现可追溯至计算机发展的早期阶段（20世纪50年代和60年代），使用数组、链表和树等一维数据结构来表示和组织非结构化数据，之后随着数据库管理系统（DBMS）的诞生，该技术得到进一步的发展，关系型数据库管理系统（RDBMS）的引入，使得非结构化数据转化为二维列表的结构化形式加以存储与管理。近年来，伴随分布式数据库、图数据库等新型数据库技术不断涌现，非结构化数据的增量和价值持续提升，数据结构化技术一直演进和创新，以适应类型多样、来源不同的非结构化数据的揭示与组织。梳理该技术的发展历程后发现，从早期以二维列表形式描述非结构化数据转变为以多种组织形式表示与关联非结构化数据，数据结构化技术难逃以二次数据揭示与组织数据资源的本质。在该技术的演变过程中，早期注重非结构化数据的描述著录，实现结构化的存储与管理；后期聚焦非结构化数据的灵活关联与组织，用以挖掘数据单元的潜在语义联系和应用价值。其中，著录、索引等对文献的内容、形式、位置特征加以揭示的目录学方法，在用属性-属性值的形式著录描述非结构化数据的过程中均有一定的借鉴；此外，数据结构化技术后期发展阶段的非关系型数据库，多根据数据的属性内容特征产生连接关系加以组织，这一点亦体现在通过内容特征关联组织文献的分类法、主题法等目录学方法中，并对目录学的文献组织方法有一定的拓展与改善。

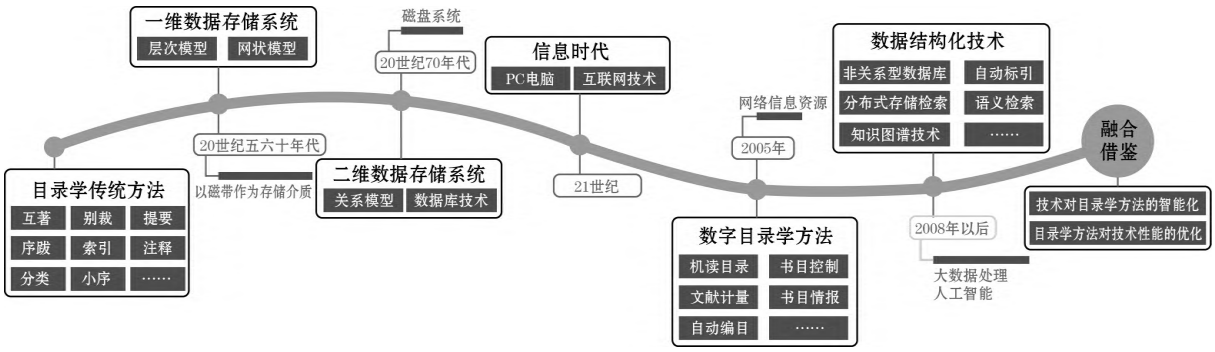

图5 目录学方法与数据结构化技术的协同演化脉络

在数据结构化技术不断发展的推动下，目录学的方法也产生了一定的变化。如：文献著录方法在引入基于机器学习的特征揭示技术后实现了自动化，并对著录特征的格式和内容提出了统一化、标准化的新要求；索引可通过语法规则或词表完成自动化的构建，依据用户需求实现个性化的排序，赋予了传统索引工具灵活性的新特点；分类法依据训练数据、分类规则构建分类模型或算法，依据资源内容特征实现基于计算机的自动分类。数据结构化技术在引入目录学方法的实施过程后，赋予了其自动化、高效率的特点，对资源语法特征的揭示与组织是行之有效的，但对于涉及内容等语义或语用特征的主题标引等目录学方法，其准确性仍有待提升。因此，目录学方法还需借助不断发展更迭的技术，才能在未来历久弥新、焕发新生、代代相承。

目前，数据结构化技术继承借鉴了多种目录学方法原理，提高了揭示、检索、组织非结构化资源的效率，但其参考的多是围绕外在形式特征和部分内容特征而展开的目录学方法，且在对待内容特征的处理上难以通过计算机达到人工操作的质量水平。此外，目录学的具体方法还包括大小序、提要、辨嫌名、辑佚书、书掌于官、广储副本以备勘改伪谬、若有所改必载原文、著采残逸以补阙漏等，这些观点具有开创性意义，代表了我国古代目录学理论研究的最高成就[53]，其中渗透的目录学思想仍有技术借鉴利用的价值，在技术的后续发展过程中仍可适当继承与完善。

目录学方法经历了由单一走向多元的多重转型，由于一贯强调致用性，在类例、小序、解题、别裁、互著等实践层面的编纂方法一般带有过于浓厚的经验描述色彩，在广泛借鉴计算机、社会和人文等其他科学领域解决问题的思维方式的基础上，将系统论方法、信息论方法、控制论方法、数理方法等引入目录学领域后，赋予了当代目录学方法更多的理性色彩，出现了诸如文献计量、书目控制、引文分析等一系列新方法[3]。但是，目录学的部分传统方法在处理文献资料时包含有语义化和语用化取向，目前的理性化技术手段代替人工完成文献在语法层次的整合时尚可，但无法如学术大家的水平整理出诸如提要、序跋等主观性内容，后续发展或可引入智能化程度更高的技术手段提高效率和准确率，或可分解其中的主观性和客观性内容，以计算机辅助参与的形式节省人力成本。

4 结语

长久以来，围绕“信息科学技术挑战目录学学科定位”的论调一直甚嚣尘上，对此，已有不少研究对此展开了辩驳与回应，但多以一种割裂的视角辨析了信息技术与目录学的关系。然而，二者的关系并非“泾渭分明”，信息技术的产生与发展过程中离不开目录学方法理论体系的支撑，目录学的传承、发展与应用也有赖于新兴技术的创新与推广，二者更多呈现一种相辅相成、互为表里的共生共存关系。本研究立足于此，剖析数据结构化过程的具体应用技术，在原理层面挖掘其中蕴涵的目录学方法原理，并进一步梳理了目录学方法论在当下技术中的发展与演变，从细节层面展现了目录学与信息技术的紧密联系，揭示数据结构化技术的底层原理过程中蕴涵的目录学方法，说明数据结构化技术是现代信息技术对目录学方法论的一种借鉴、继承、发扬与完善。

目录学作为应用、辅助性的学科，因其致用价值而被其他学科广泛需要，常于无形之中将其方法理论渗透融入其他学科之中，例如本文揭示的数据结构化技术中的目录学方法论，即是目录学向计算机科学领域的延伸与拓展。其中，信息科学技术作为目录学发展的新机遇，揭示其与目录学的关联关系，是当前探寻与发展目录学理论新的突破口。阐明数据结构化技术中运用的目录学方法论，是彰显目录学致用属性的有效途径，也是目录学在信息技术兴起背景下应有的开放包容视角，助力目录学在技术迅猛发展的时代下既能守正传承思想方法，又可与时俱进地出陈纳新，在未来的传承与发展过程中焕发新的生命力。

参考文献

[1] 费巍．西方目录学的发展及其对我国目录学研究的借鉴意义 [J]．图书情报知识，2008（1）：50-57+104.

[2] 刘培旺．新文科建设与目录学学科发展——第七届全国目录学学术研讨会综述 [EB/OL]．[2023-08-23]．https://kns.cnki.net/kcms/detail/42.1085.G2.20230704.1015.002.html.

[3] 柯平，刘旭青．中国目录学七十年：发展回溯与评析 [J]．中国图书馆学报，2019（5）：101-111.

[4] Voorsluys W，Broberg J，Buyya R. Introduction to Cloud Computing [M]．Hoboken：John Wiley & Sons，2011：3-42.

[5] Brown T B，Mann B，Ryder N，et al. Language Models Are Few-Shot Learners [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Canada，2020：1877-1901.

[6] Nakamoto S. Bitcoin：A Peer-to-peer Electronic Cash System [EB/OL]．[2023-08-23]．https://bitcoin.org/bitcoin.pdf.

[7] 石进，胡雅萍，李益婷．大数据时代目录学的新使命 [J]．图书馆学研究，2019（6）：49-55.

[8] 熊翔宇，郑建明．大数据管理中的目录学思想 [J]．图书馆学研究，2019（12）：2-8.

[9] 彭贤哲，郑建明，李佳新，等．目录学思想在数据结构化过程的传承与应用 [EB/OL]．[2023-08-23]．https://kns.cnki.net/kcms/detail/42.1085.g2.20230711.1030.002.html.

[10] 马梦丹．论古典目录学与现代目录学的异同及目录学的发展趋势 [J]．科技文献信息管理，2019（4）：9-14.

[11] 张琦，邵彦敏．智慧校园背景下信息协同评价与推进策略研究 [J]．情报科学，2019（8）：102-107.

[12] 王锰，陈雅，郑建明．国内外数字时代的目录学理论体系研究进展 [J]．图书馆，2014（6）：32-37.

[13] 章成志，谢雨欣，宋云天．学术文本中细粒度知识实体的关联分析 [J]．图书馆论坛，2021（3）：12-20.

[14] Chu H T，Ke Q. Research Methods：What's in the Name? [J]．Library & Information Science Research，2017，39（4）：284-294.

[15] Guo Y，Silins I，Stenius U，et al. Active Learning-Based Information Structure Analysis of Full Scientific Articles and Two Applications for Biomedical Literature Review [J]．Bioinformatics，2013，29（11）：1440-1447.

[16] Duck G，Kovacevic A，Robertson D L，et al. Ambiguity and Variability of Database and Software Names in Bioinformatics [J]．Journal of Biomedical Semantics，2015，6（1）：29.

[17] Lin L，Wang D，Shen S. Extraction of Thesis Research Conclusion Sentences in Academic Literature [C]//EEKE2021-2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents. Online，2021：74-76.

[18] Mesbah S，Lofi C，Valle T M，et al. TSE-NER：An Iterative Approach for Long-Tail Entity Extraction in Scientific Publications [C]//International Workshop on the Semantic Web. Porto，2018：127-143.

[19] 商宪丽，王学东，张煜轩．基于标签共现的学术博客知识资源聚合研究 [J]．情报科学，2016（5）：125-129.
[20] 曾建勋．中文知识链接门户的构筑 [J]．情报学报，2006（1）：63-69.
[21] 成全，周兰芳．面向语义关联的微博信息多维主题聚合研究 [J]．情报理论与实践，2018（7）：136-142.
[22] Coussement K，Benoit D F，Antioco M. A Bayesian Approach for Incorporating Expert Opinions into Decision Support Systems：A Case Study of Online Consumer-Satisfaction Detection [J]. Decision Support Systems，2015，79：24-32.
[23] Gadomer L，Sosnowski Z A. Knowledge Aggregation in Decision-Making Process with C-Fuzzy Random Forest Using OWA Operators [J]. Soft Computing，2019，23（11）：3741-3755.
[24] Zheng C，Chen Y，Chen C，et al. Density - Based Clustering with Kernel Diffusion [EB/OL]. [2021-10-01]. https://ui. adsabs. harvard. edu/abs/2021arXiv211005096Z.
[25] 黄文碧．基于元数据关联的馆藏资源聚合研究 [J]．情报理论与实践，2015（4）：74-79.
[26] 宋爽，张国栋．国内外同构聚合检索系统比较研究 [J]．大学图书馆学报，2011（5）：55-59.
[27] 许海云，董坤，隗玲，等．科学计量中多源数据融合方法研究述评 [J]．情报学报，2018（3）：318-328.
[28] 王萍，牟冬梅，杨鑫禹，等．基于数据特征的在线健康社区信息融合模式研究 [J]．现代情报，2022（8）：28-36+167.
[29] Oualhi O L，Mohamed T. Dynamic Generation of Adaptative Teaching Material for Semantic Web Approach [C] //IADIS Multi Conference on Computer Science and Information Systems. Lisbon，2012：7-11.
[30] 彭斐章，陈传夫．目录学教程 [M]．北京：高等教育出版社，2004：205.
[31] Berners-Lee T，Fischetti M，Dertouzos M L. Weaving the Web：The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor [M]. New York：Harper San Francisco，1999：177-198.
[32] Shadbolt N，Berners-Lee T，Hall W. The Semantic Web Revisited [J]. IEEE Intelligent Systems，2006，21（3）：96-101.
[33] 谢力军，杨军．几种索引技术的比较 [J]．怀化学院学报，2009（8）：115-118.
[34] 陈艳．信息检索可视化技术 [J]．情报理论与实践，2006（5）：618-621+566.
[35] 韩红旗，桂婕，张运良，等．大规模主题词自动标引方法 [J]．情报学报，2022（5）：475-485.
[36] 刘熠，张智雄，王宇飞，等．基于语步识别的科技文献结构化自动综合工具构建 [EB/OL]. [2023-08-28]. http://kns. cnki. net/kcms/detail/10. 1478. G2. 20230504. 1651. 002. html.
[37] Xun G，Jha K，Zhang A. MeSHProbeNet-P：Improving Large-Scale MeSH Indexing with Personalizable MeSH Probes [J]. ACM Transactions on Knowledge Discovery from Data，2020，15（1）：1-14.
[38] You R，Liu Y，Mamitsuka H，et al. BERTMeSH：Deep Contextual Representation Learning for Large-Scale High-performance MeSH Indexing with Full Text [J]. Bioinformatics，2021，37（5）：684-692.
[39] Sheth A，Padhee S，Gyrard A. Knowledge Graphs and Knowledge Networks：The Story in Brief [J]. IEEE Internet Computing，2019，23（4）：67-75.
[40] 夏翠娟．知识组织方法和技术的演变及应用 [J]．晋图学刊，2021（6）：1-9.
[41] 朱良兵，纪希禹．基于 Topic Maps 的叙词表再工程 [J]．现代图书情报技术，2006（9）：81-84.
[42] 张玉涛，夏立新．基于主题图的电子政务信息资源整合模型研究 [J]．情报杂志，2009（7）：161-165.
[43] 施旖，熊回香，陆颖颖．基于主题图的非物质文化遗产数字资源整合实证分析 [J]．图书情报工作，2018（7）：104-110.
[44] 万长松．MMK：苏联（俄罗斯）科学哲学的摇篮 [J]．自然辩证法通讯，2023（8）：1-10.
[45] 郑建明．当代目录学 [M]．北京：科学出版社，2020：76.
[46] 柯平，张颖．呼唤 21 世纪的新目录学——柯平教授访谈录 [J]．图书馆，2022（10）：1-7.
[47] 马芝蓓．当代目录学方法论体系探讨 [J]．图书情报工作，1994（2）：24-29+47.
[48] 安小米，王丽丽．大数据治理体系构建方法论框架研究 [J]．图书情报工作，2019（24）：43-51.
[49] 胡唐明，郑建明，黄建年．衔接与融合：当代目录学研究进路 [J]．图书馆理论与实践，2012（2）：4-8.
[50] 彭斐章．目录学是读书治学的必修之学 [J]．图书情报知识，2014（3）：8-9.
[51] 曾伟忠．试论目录学的动力因素和发展方向 [J]．图书馆学研究，2010（5）：2-6+28.
[52] 潘文年．清代中前期的民间刻书及其文化贡献 [J]．安徽大学学报（哲学社会科学版），2008（2）：142-148.

彭贤哲　南京大学信息管理学院博士研究生。研究方向：目录学、大数据分析与技术。
郑建明　南京大学信息管理学院教授，博士生导师，博士。研究方向：数字信息资源管理、目录学基础理论。
石　进　南京大学信息管理学院教授，博士生导师，博士。研究方向：智能目录学、大数据分析。通讯作者。