基于数据要素流通视角的数据溯源研究进展*

王晓庆[1,2,3]　孙战伟[1]　吴军红[4]　杜自然[5]　钱城江[6]

[1]（南京财经大学公共管理学院 南京 210023）

[2]（南京航空航天大学经济与管理学院 南京 211106）

[3]（南京财经大学红山学院 南京 210003）

[4]（南京师范大学商学院 南京 210023）

[5]（深圳市数聚湾区大数据研究院平台研发部 深圳 518048）

[6]（南京南工大安全科技有限公司 南京 210047）

摘要：【目的】通过文献梳理分析数据溯源研究进展及应用场景，以期为数据交易平台搭建、行业数据治理建设和数字政府治理建设提供参考。【方法】从数据溯源模型、数据溯源方法和数据溯源应用分别进行归纳和分析，并在此基础上探讨研究现状和不足之处。【结果】无论是在内容描述、模型构建，还是场景应用方面，数据溯源研究均取得了丰富成果，表现为数据溯源质量得以提高、数据溯源安全得以保障、数据溯源效率得以提升。【局限】基于要素流通视角对数据溯源的研究起步相对较晚、研究成果不够丰富、研究体系尚未形成、研究重点偏向实证。【结论】可从与数据要素市场相结合，积极推进数据交付使用常态化；加快推进数据溯源标准工作，积极推进数据使用工作制度化；不断提升数据溯源信息质量，积极推进数据服务优质化；高度重视数据溯源信息安全，积极推进数据信息使用规范化；高标准搭建数据溯源平台，积极推动数据要素市场健康化发展等方面进行深入研究。

关键词：数据流通　数据溯源　管理模型　数据要素

分类号：TP391

DOI：10.11925/infotech.2096-3467.2022.0017

引用本文：王晓庆，孙战伟，吴军红等. 基于数据要素流通视角的数据溯源研究进展[J]. 数据分析与知识发现，2022，6(1)：43-54.(Wang Xiaoqing, Sun Zhanwei, Wu Junhong,et al. Research Progress of Data Traceability from the Perspective of Data Element Circulation[J]. Data Analysis and Knowledge Discovery, 2022, 6(1): 43-54.)

1　引　言

2020年4月，中共中央、国务院在《关于构建更加完善的要素市场化配置体制机制的意见》中明确提出“加快培育数据要素市场”，强调数据作为要素的重要意义。2021年12月，中国信息通信研究院发布《大数据白皮书》，亦将“激活数据要素潜能、加快数据要素市场化建设”作为核心议题。那么，何为数据？数据从狭义来讲就是数字、数值，广义上数据也可以是文字、图像、声音等。而信息化时代又诞生出“大数据”概念，大数据具有规模巨大、类型多样、传播速度快、传播渠道多且无处不在、无时不有等诸多特点，已经渗透到政治、经济、军事、文化等多个重要领域，成为经济发展、社会稳定、国家安全不可或缺的重要资源。当然，这些重要数据不是“天生”的，而是在特定设备中经过特定程序筛选出来的[1]，任何

通讯作者（Corresponding author）：钱城江（Qian Chengjiang），ORCID：0000-0002-0559-005X，E-mail：qiancj_njtech@163.com。

*本文系国家社会科学基金青年项目（项目编号：18CSH018）的研究成果之一。

The work is supported by the National Social Science Fund of China (Grant No. 18CSH018).

不明来源的数据都是没有任何价值的、也不应该被运用到重要场景中，数据来源及数据安全的重要性愈发凸显。人们通常所使用的数据并非原生数据，而是经过数据要素流通市场并依照相应规则、标准进行加工、计算、聚合、交易后的派生数据。从本质上讲，原生数据“杂乱无章”，不宜直接用于特定的重要场景；经过“深加工”形成的派生数据更适合用于特定的重要场景，但须保证派生数据的真实性、可靠性和安全性。原生数据“深加工”后已变得“面目全非”，为重现原生数据的真实状态，就必须采用特定的方法或路径去回溯，即为数据溯源。

随着新一代信息技术的迅猛发展，数据已经成为重要的资源，发挥数据价值的关键在于数据流通。数据流通指某些信息系统中存储的数据作为流通对象，按照一定规则的从供应方传递到需求方的过程，主要以一对一许可、互为许可和一对多（众）许可方式进行流通。数据要素的流通是创造数据价值的关键一环，数据流通可以是简单的阅读或识读，也可以是对大量数据进行演算分析。基于数据演算分析的一对一许可数据流通过程如图1所示。

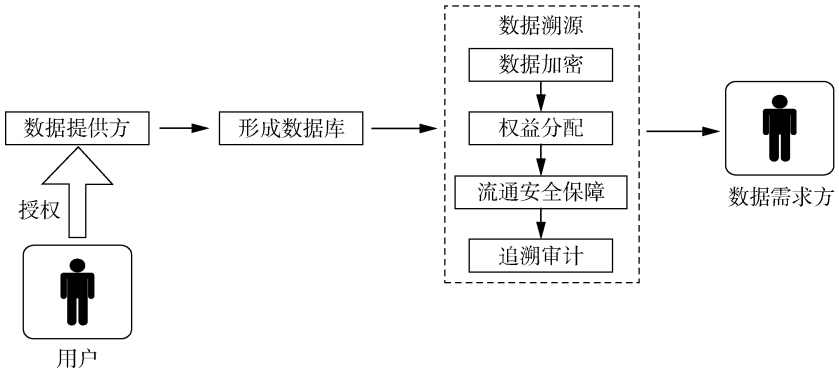

图1 数据流通图

Fig.1 Data Flow Process Diagram

随着数字经济逐渐进入高质量发展新阶段，数据要素市场对数据的安全治理、安全保护提出更高要求，构建可追溯的数据要素市场保障数据要素流通安全、规范被提上日程，数据溯源在数据治理中的重要性也愈发凸显。基于数据溯源的重要性，以“数据溯源”为主题词在知网进行检索，获取的相关文献关键词聚类分析结果如图2所示，聚类中心关键词分别是数据溯源、区块链、大数据、隐私保护等。对上述文献进行回溯与归类，发现较多集中在数据溯源模型、数据溯源方法、数据溯源应用等三类主题上，因此，本文在对相关文献进行系统性梳理的基础上，对于数据溯源模型、数据溯源方法及数据溯源应用等核心议题进行总结归纳，由此揭示当前数据溯源研究的方向，进而对未来研究进行展望。

2 数据溯源模型

数据要素通过要素市场流通、通过数据交易平台进行交易，需要明确的交易规则、完备的交易制度，即数据要素包括哪些内容、会有什么价值、怎样衡量价值、如何实现价值，需要借助数据管理模型表现。数据管理模型承载着数据需求的元数据集合，是数据质量校验的对象、数据集成与互操作的起点、数据存储和操作的结构保障、数据仓库和商业智能（Business Intelligence，BI）的核心[2]。由此可见，在数据溯源体系中，数据管理模型起着举足轻重的作用。只有构建好数据溯源管理模型，才能明确数据溯源的技术路径与操作步骤。当前研究文献和工作实践中，“出镜率”较高、“知名度”较响、影响力较强的当属以下三种模型。

2.1 数据溯源描述模型

数据溯源描述模型主要功能在于记录数据溯源信息，对这个问题的研究经历了多个阶段。早期对数据溯源信息的研究，较多关注数据从哪里来，通常做法是将数据及其有关信息共同存储以备查询使用。随着对数据溯源信息使用范围的扩大和学术研究的不断深入，数据溯源描述模型从W2（Why、

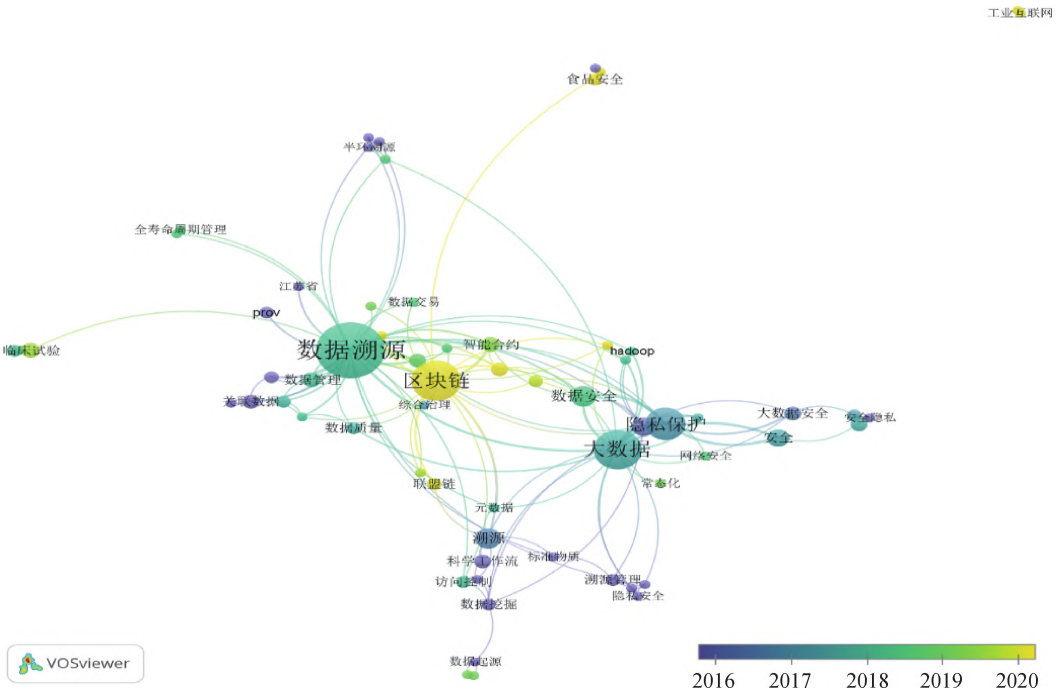

图 2　文献检索结果聚类分析

Fig.2　Cluster Analysis of Literature Retrieval Results

Where）模型[3]，扩展到 W3（Why、Where、How）模型[4]，再到如今较为成熟的 W7（Who、Why、Where、How、What、When、Which）模型[5]。与 W2 模型、W3 模型相比，W7 模型的进步之处在于，它不仅对数据溯源的范围进行了拓展，而且明确了数据溯源的核心要素 What，即在数据的全生命周期中，数据发生了什么、应该是什么，形成了基于工作流的数据溯源模型。

然而，W7 模型也存在固有缺陷，它无法全部囊括和涵盖科研数据管理全过程，对科研数据的溯源手段不够多元、溯源内容不够精细。于是在 W7 模型的基础上，衍生出覆盖内容更为全面、溯源内容更为精细的“W7+R3”科学数据溯源内容模型[6]。科学数据溯源内容模型的进步之处在于，不仅丰富了 W7 模型内容，增添了事件资料、事件结果和事件标注等模型元素，更重要的是，模型能够有效记录数据密集型科研项目不同阶段的重要事件及关键信息，进一步规范科学数据溯源中的元数据，进一步细化模型中各个要素之间的映射关系，能够使得科研过程的可复制、可再现、可验证得以有效实现。

无论是 W3 模型，还是“W7+R3”科学数据溯源内容模型，都不能解决不同系统之间溯源信息的交换。基于此类原因，在 2007 年的溯源竞坛工作组会议上，部分学者研究并发布了开放溯源模型[7]（Open Provenance Model，OPM）。OPM 模型从技术角度实现了不同系统之间交换时可用信息的溯源，允许并支持不同层级的描述同时存在，逐渐成为应用领域最为广泛的数据溯源信息交换的执行标准。随着 OPM 模型应用的逐步深入，数据溯源信息的研究呈现出复杂化、多样化、特色化等特点，针对不用应用情景且由不同行业组织推出的数据溯源模型也日益多元。能够定义一种通用的溯源模型和表达语言、规范不同溯源模型之间统一交换的 PROV（Provenance）溯源标准由此产生，成为数据溯源史上的里程碑。此后，围绕 PROV 溯源标准，不同组织、不同学者对其进行深入更深层次的研究[8-11]，在 PROV 溯源标准的基础上，探讨其在不同应用情景下的适用范围，并与 OPM 模型进行对比研究；推出面向用户、开发人员、管理人员等不同使用对象的说明文档，并开展相关培训，进一步将对其的重视程度推向高潮。

为规范数据采集、发布、分析和处理，2017 年国

家质量监督检验检疫总局、国家标准化管理委员会在联合发布的《信息技术数据溯源描述模型》中，定义了ProVOC（Provenance Vocabulary Model）数据溯源描述模型[12]，该模型由数据、活动和执行等实体组成，是根据数据发展实际情况、解决数据交易难题大背景下对PROV溯源模型进行的改进。

2.2 数据溯源应用模型

随着溯源技术的发展，数据溯源模型不断得以扩展并在特定领域得到广泛应用，将Data、Process、Agent作为主要组件的Provenir数据溯源应用模型[13]诞生。其中，Data代表用于科学实验的材料、产品及相关参数，而Process、Agent与开放溯源模型中所指涵义具有相似之处。但与之具有明显不同，在Provenir数据溯源模型中，Process被认为具有随外在因素变化而发生变化的特性，属于Occurrent；Data、Agent则不具有随外在因素变化而发生变化的特性，属于Continuant。

同时，基于方便查询特定数据产品来源、合理评估特定数据产品质量以及重现特定数据产品的产生过程，在使用SAR数据分析方法的处理基础上构建出基于分层二部图的溯源模型[14]。值得关注的是，通过对江苏省地面沉降监测系统进行实证研究，能够充分说明上述模型不仅适合处理SAR数据的溯源，也适合处理复杂化情况下高密度数据的溯源。为应对空间数据溯源缺乏标准的棘手问题，有学者提出包括空间数据溯源服务、关联空间数据溯源RDF（Resource Description Framework）数据库、关联自动构建中间件、客户端和本体知识库等部件的空间数据溯源模型，用以解决空间数据溯源信息的存储与发布机制及智能搜索的高效、便利等问题[15]。

此外，针对特定行业、具体领域的数据溯源应用模型中，还有主要应用于医疗行业的TVC（Time-Value Centric）溯源模型、主要应用在地球学领域的四维起源模型，以及主要适用于光学系统的流起源信息模型等。

2.3 数据溯源安全模型

同数据本身一样，数据溯源对安全也有极高的要求，因为面对不可信的复杂环境，经历不同的处理流程，数据溯源信息面临着被编辑、被转换、被篡改等层层失真的安全风险。通常情况下，重新构建数据溯源模型，或改进优化原有数据溯源模型，建立新的信任机制或采用安全方法，也能有效检测和甄别溯源链条中数据信息的真实性与完整性。如根据数据溯源的安全需求及广播加密方案，构建新的数据溯源安全模型，这样即可采用密钥树再生长方法解决机密性需求中审计用户数量动态变化问题[16-17]。除建立全新的数据溯源安全模型外，也有部分学者不断对已有的数据溯源模型进行扩展，如对PROV模型进行扩展，构建基于医疗健康大数据的安全溯源模型[18]；将各种数据溯源模型合理组合，构建涵盖安全层、逻辑层、语义层等分层数据的溯源安全模型[19]。

数据溯源安全与区块链技术紧密相连、密切相关。区块链技术，也称为“共识技术”，因对存储于其中的数据或信息具有不可伪造、全程留痕、可以追溯等特点受到学者青睐，符合数据溯源安全管理的特性与要求。据此，构建基于区块链技术的溯源管理模型成为数据溯源安全模型的可能选择，也是重要选择。一种策略是，将区块链技术与射频识别（Radio Frequency Identification，RFID）技术有机结合，构建起多个主体参与、多个部门协同、信息公开透明、数据真实共享的溯源链条，以及涵盖溯源物品生产、销售、流通、加工、消费等溯源路径，做到RFID数据溯源安全管理[20]。另外一种策略是，将区块链技术与智能合约协议有机结合构建合约模型，探讨溯源模型在不同条件下实现的可能性、扩展性，在可信任平台中，只有诚实参与者才能够安全收集和验证溯源信息。还有的一种可能策略是，将区块链技术与云技术有机结合，不仅可以实时溯源云端数据信息，而且隐私保护能力和安全性能系数更高[21]。

当前，在数据要素流通市场上，数据的价值尚未完全得到确认，其中的重要原因就在于数据要素的真实性、合法性无法得以保障。通过数据溯源模型，呈现数据资源在不同环节、不同层次的要素价值与真实属性，为清晰界定数据权属、避免数据交易纠纷打下坚实基础。

3 数据溯源方法

所谓模型，是指为深入研究某种现象或解决某种问题，对影响其变化或导致其结果的各类关联因素及其相互关系构建的机制。方法，就是根据具体

的模型，所采用阐释现象或解决问题的有效方法。因此，数据溯源模型与数据溯源方法具有很强的逻辑关联。在要素市场上，数据一旦被交易出去，其管理主体将由一方逐渐扩散至多方，其主要内容也可能会由“真”变“假”。尤其当数据交易出现争议，甚至法律风险时，如何采用数据溯源方法识别数据在某个时点的真实性、权属性？数据溯源方法的有效构建与合理选择就显得比较重要。当前文献对数据溯源方法的研究主要集中在以下方面[22]。

3.1　面向关系数据库的溯源方法

溯源信息以不同粒度的形式存储在关系数据库中，在简单的应用场景下，通过成熟的溯源系统可以较好地进行数据溯源。事实上，许多应用场景都带有复杂的查询语句，以细粒度形式的溯源信息标注往往会产生大容量存储，并且降低溯源效率。如何对关系数据库下的溯源计算方法进行优化，成为亟待破解的难题。

作为特殊目的的数据库查询语言和程序设计语言，结构化查询语言在数据溯源中起着重要作用，但结构化查询语言不同于一般的数据库查询语言，在传统关系数据库下无法直接借助其实现数据溯源功能。通常可用策略是，面临复杂的结构化查询语言时，借助某种关联将数据溯源过程转化，对溯源数据进行计算、对溯源结果进行查询。亦有学者构建基于结构化查询语言的数据转换图形，给出模式级数据的处理过程、源表字段与目标表字段之间的映射关系，实现了模式级、字段级以及元组级三个层次的数据起源功能，包括起源解析、起源存储、起源查询和可视化过程[23]。但数据溯源效率一直是数据溯源面临的棘手难题，数据溯源效率问题的实质在于溯源过程中对数据库存储空间和存储效率的优化，因此学者们尝试设计专用查询中间件，或采用查询数来减少查询节点，筛选重要信息，降低冗余存储[24-26]。

3.2　面向科学工作流的溯源方法

科学项目最优解的计算往往呈现超高强度计算、超强复杂依赖特点，科学工作流正是为解决这种大数据高密集型科学实验而开发设计的，使用科学工作流有助于对科学研究过程中产生的相关数据进行计算、分析。顾名思义，面向科学工作流的溯源计算方法即是对不同阶段的科学工作过程及产品信息进行溯源。

面向科学数据溯源计算过程中产生的海量信息存储及计算效率问题，可以采用继承方法减少存储数量，用分解算法或更加高效的溯源方法对符合规范的科学工作流进一步优化和提升科学数据溯源效率[27]；也可以代理对象数据库中双向指针表查询为基础原型，构建双向指针溯源方法对科学任务的执行过程进行溯源，从而实现科学数据的高效追踪[28]。针对具体应用领域的科学工作流数据溯源方法，部分学者对此进行了积极探索。如在天文研究领域[29]，以 SOA（Service-Oriented Architecture）数据溯源收集框架为基础，构建基于 XML（eXtensible Markup Language）的数据溯源模型，对数据溯源的收集、存储、查询等功能进行测试。也有学者以具体项目工作流为实证案例，指出溯源信息在支持数据发布时存在的不足，并提出面向数据发布的科学工作流溯源方法[30]。考虑到分布式账本技术在交易环节发挥的重要功能，通过与科学工作流系统共享数据溯源信息，探讨分布式账本技术与科学工作流在加密算法、共识共享等方面的异同点，从而探究可行方法[31]。另有学者设计了面向科学应用的分层溯源采集和查询框架，主要包括溯源数据收集层、溯源数据存储层、溯源数据分析层[32]。该模型实现了收集存储和分析存储相分离，能够灵活支持多种溯源模型，可扩展性比较高。现有科学工作流下的溯源方法研究更多侧重于数据溯源效率，未来需要在关注数据溯源效率的同时注重数据溯源安全的方法研究。

3.3　面向大数据平台的溯源方法

开展数据溯源研究，能够有效降低和排除大数据时代、大数据平台下跨领域、多源异构、数据不断转换和变化带来的失真、失信风险。设计 MapReduce 的初衷是为了解决搜索引擎中大规模网页数据的并行化处理，主要程序包括 Map 函数、Reduce 函数和 Main 函数。2004 年，Doug Cutting 在 MapReduce 的基础上进行技术升级和算法迭代，推出可靠性好、扩展性强、处理能力高的 Hadoop 系统。此后，将 RAMP 作为 Hadoop 的溯源组件用以收集 MapReduce 工作流信息，同时对 Map 函数、Reduce 函数加装包装器，减少人工对其干预程度[33]。当然，

对于数据溯源过程中发现的大数据错误，可以采用Newt工具进行有效纠正。

随着云计算服务的运用和推广，云计算具备的数据处理高效、数据存储量大、数据稳定性强等特点不断显现，因此云计算环境下进行数据溯源可以提供更为完整的过程。建立通信通道对虚拟层和物理层进行统一管理，实现云环境下存储虚拟化，能够快捷、安全地对数据溯源信息进行访问。还有学者试图根据云服务各类节点产生的日志数据建立基于日志的云服务溯源方法，如Flogger、S2Logger和Progger。其中，Flogger具有很强的可扩展性，能够增强高性能日志记录的可读性；Progger可用于追踪文件的数据活动[34]。或借助云环境背景重新构建新颖的数据溯源模型，对数据溯源的记录、搜集、存储与查询等方法进行分析与探讨[32]。

3.4 面向区块链的溯源方法

只有不断优化区块链技术框架中的共识机制，才能使得基于云计算与区块链双重技术的加持下的溯源效率更高；只有全方位使用区块链技术背景下应用程序对溯源信息的威胁内容，提出相应的安全溯源模型才会更具有可操作性。经过相应的程序测试和评估表明，安全溯源模型能够对区块链应用下的溯源数据起到很好的保护作用[35]。区块链应用下的溯源计算模型，对工作量证明、权益证明的共识机制优化内容进行研究。权益持有人越维护网络利益，就会得到越多的激励，则越有机会做更多的决定[36]。

共识机制的不断优化，能够解决分布式场景下平衡系统的性能效率和资源消耗的均衡问题，智能合约同样需要解决一致性问题。不同的是，智能合约是通过执行任务合约的方式扩展区块链对数据的处理能力，这样在区块链中保存处理数据的同时也对智能合约进行保存，从而实现各个节点下智能合约的一致性。基于区块链中的智能合约具有不可变更性特点，使得用户信赖区域链的同时，也增加了区块链在不同场景中的应用可能。对形式化本体概念、性质以及内容进行分析，并以产品供应链溯源为实证对象，对多伦多虚拟企业溯源本体方法与智能合约的转换过程进行研究，验证了该智能合约可以对数据进行溯源及通过区块链进行溯源行为约束[37]。

对数据的可靠性和高质量要求加快了区块链技术的发展，但现阶段对基于区块链技术开展的数据溯源还处于探索阶段，较多集中在概念层面，少量集中在特定领域和应用行业。区块链涉及专业知识的多重性、技术领域的复杂性，展现出对使用区块链技术如何合理存储原始数据及标注信息、如何描述溯源记录并确保可信的研究内容较少。但可以明确的是，借助面向区块链技术的数据溯源等方法，将会为数据要素在流通时提供可靠的技术支撑，有利于激活沉淀在个人手中、企业内部、政府平台的数据资源，推动数据要素由资源向资产转化。

4 数据溯源应用

数据安全有序流动，有助于帮助企业更好地掌握市场需求、推进产品迭代升级；而数据跨境流动与交易带来的全球价值功能，更能凸显国家的国际话语权地位。因此，数据溯源模型与数据溯源方法的主要价值体现在事关国计民生的具体场景应用上。根据当前研究，数据溯源应用主要集中在突发重大事件、电子商务领域、企业经营以及科学研究等场景。

4.1 突发重大事件应用场景

SARS事件、汶川地震、南方雪灾、烟台笏山金矿爆炸、全球新冠病毒等突发重大公共事件中，既有自然灾害，又有事故灾难，短期内对人民生命和公共财产造成严重损害。面对广为传播的谣言等各类舆情，如果社会救治信息不够透明、调查信息公布不够及时，无疑会对受害者及其家属的心理产生二次伤害。全媒体时代下，政府如何利用大数据技术及时、全面掌握突发重大公共事件的真实信息，如何短期内采取有效措施应对与处理突发重大公共事件网络舆情与社会谣言，考验着政府管理者的智慧与能力。在以往突发事件的处理中，较多采用网络舆情监测软件通过搜索、识别关键词等方式对网络舆情进行管控。这种方式虽然比传统的人工筛选、上报机制在时效性上有所改观，但对网络舆情的精准识别还存在不足之处。

很明显，对重大突发事件的有效处置，离不开对相关信息的数据溯源、模型建立及结果预测。在流

感监测研究中，将一元线性回归模型、多元线性回归模型、主成分回归模型和人工神经网络模型在流感监测的使用情况进行分析与对比，引入官方对流感监测的数据推导的优化模型表明：历史数据和搜索数据的真实性、完整性是预测结果准确的重要前提[38]。采用区块链溯源技术对重大公共卫生事件的原始数据进行溯源，分析数据来源信息与修改过程，可以有效对重大公共卫生事件做出舆情研判[39]。平安科技与重庆疾控中心联合课题组依托医疗数据，采用人工智能、大数据等技术，构建并通过模拟验证测试的基于AI的流感预测模型[40]，更是有力地改变了发达国家在公共卫生事件上预防为主的霸主情形，使得中国从对公共卫生事件的事中事后处理逐步转向以事前预防为主。

此外，产生重大突发事件并造成负面舆情后，如何快速、准确地进行数据溯源？以“美联航拖拽亚裔”突发事件作为实证对象，研究突发事件的信息瀑布溯源机制，构建基于时间序列与信息融合的突发事件信息瀑布溯源模型，在丰富和完善信息溯源模型的同时，有效助力政府部门、非政府专业组织对突发事件的信息瀑布进行有效溯源[41]。微博作为当下广泛使用的传播媒介，使用过程中不可避免地存在垃圾用户、负面信息等，一旦形成舆情，极易在段时间内快速传播扩大并造成重大突发事件。构建基于MACD（Moving Average Convergence and Divergence）的突发事件检测算法[42]，通过过滤垃圾信息、切分微博内容等预处理方式，对突发事件进行聚类形成事件集合，从而采取有效应对措施将突发事件带来的影响降至最低。

4.2　电子商务应用场景

执行法律法规与公共政策，遵守安全协议与技术标准，电子商务系统实现了产品生产者和消费者在线交易，推动了网络经济大力发展。然而，采用电子商务系统也存在诸多痛点难点，如虚拟网络、跨地区、跨部门、监管体制、配套法律等多种因素相互交织带来的电子商务系统中交易的安全问题；食品通过电子商务系统进行存储、运输、配送环节，容易出现数据链条监管缺位导致食品中毒、过期等安全问题。

我国是世界上电子商务发展较为快速的国家，因此对电子商务领域的监管格外关注。为有效防止不法人员恶意篡改电子商务系统数据，政府部门、行业协会、大型商贸流通企业纷纷搭建食品安全溯源平台。政府层面，工业和信息化部启动的食品工业企业质量安全追溯平台，先期开放乳粉产品追溯数据，最终达到监管行业产品信息的即时跟踪。行业方面，非营利性行业组织——中国副食流通协会食品安全与信息追溯分会搭建的中国食品安全信息追溯平台，为食品企业提供第三方信息追溯服务和数据交换平台服务。企业方面，京东采用区块链技术推出“京东万象”数据溯源平台，对交易数据实时加密管理，确保交易数据安全可靠[43]。天猫国际综合运用区块链技术，将商品从生产环节到流通环节的所有信息嵌入到商品的编码中，消费者通过扫码即可实时查询商品安全状况。

4.3　企业经营应用场景

随着信息技术的演进与发展，企业面临着愈加复杂的信息环境，计算方法的优化、原生数据的质量对企业制定绩效指标、发展指标的真实性、可靠性，以及企业的管理和决策都有着重要影响。企业发展事关国计民生，因此对各类企业的数据管理和指标管理应该更为科学、规范和精细。

指标数据的溯源应能完整记录从原生数据转换为管理指标的过程。W7模型虽能作为数据指标溯源的基础，但考虑到其精细度和完整性，有必要先对W7模型中的Who和How进行适度扩展用以解决责任认定问题，并对Index和Result进行适度扩展以解决层级指标的重要性问题。同时，为规范描述指标溯源关系路径，使用OPM溯源表达模型的3种基本要素和5类关联关系为基本指引，重新定义了3类要素之间的4种关系（被执行关系、被触发关系、输入关系、输出关系），构建轻量化的指标数据溯源表达模型。使用此种指标数据溯源范式勾勒的企业经营指标，能够清晰表达出各个层级、各个地区的指标及原生销售数据的溯源过程[44]。

4.4　科学研究应用场景

在科学技术创新活动中，重要科学数据的存储与分享是极为重要的因素。确保科学数据的真实性、准确性与科学性是研究人员进行科学实验的基本前提和重要基础。对科学数据进行溯源，意在保

证科学试验的数据可以被真实记录、有效存储。区块链所具有的可追溯、防修改特性正好可以使其与科学数据有机结合。使用数据标识技术能够对溯源链条上的科学数据迅速实现精准定位，科学数据的溯源信息一旦被存储到区块链上，其他科研人员即可通过安全途径有效获取和实现共享[45]。

同时，建立共享平台并向特定人员开放科学数据已成为当今国际通行做法。英国发布《制定数据管理与共享计划》，阐述了制定数据管理与共享计划的意义与基本方法，认为管理与共享数据可以带来保持工作延续性、避免不必要的数据重复采集、保存支持文献的数据、开展更多合作、提高研究的显示度等诸多益处[46]。中国科学院计算机网络信息中心牵头承担的科学数据共享服务平台，建成了物理上分布、逻辑上统一的数据共享服务环境，实现了分布式数据资源的整合共享服务，为百余项重大科研工程、项目提供重要数据支撑服务，拥有超过百万、遍布全国且覆盖我国基础科学研究与应用领域的稳定用户群体。

类似的重要科学平台陆续建立，或已经建立正在考虑开放，但对平台的使用结果评价却不够理想。究其原因在于，哪个责任主体对共享数据的入口进行把关，哪个责任主体对共享数据的使用质量进行评价。通过科学创建数据共享生态环境，创造性融入数据溯源技术，将其与科学数据过程监管有机结合起来，通过区块链来实现科学数据的交互共享机制势在必行。如通过改造已有数据溯源管理软件，构建一套符合人文社会科学特点的数据共享平台用以解决数据共享难题，正是数据溯源技术与方法在科研数据共享领域的大胆尝试[47]。

4.5 溯源场景应用分析

数据溯源技术不管是在重大突发事件应对、电子商务系统发展，还是企业经营指标科学细化、科学研究数据的存储与共享应用上，利用区块链、人工智能、大数据等技术，既能通过数据溯源技术找到数据源头进行管控，也能找到存储的原始数据进行共享。深海空天、天文水利、生物医药、量子信息、基因技术等产业方向需要跨部门、高效率协同，产生的高密集型数据及数据处理要求愈发凸显数据溯源技术的重要。此外，我国各地利用信息技术纷纷打造数字政府，其中较为典型的模式包括：广东模式、浙江模式和贵州模式。不管何种模式，数字政府的核心都是采用区块链技术的数据驱动，如将区块链技术应用于政府统计调查领域，确保数据能够安全追溯；将区块链技术运用到政务数据领域，确保政务数据信息安全共享等。

各领域数据溯源技术应用如表1所示。然而，目前缺乏在复杂条件下，对这些高精尖行业高密集数据溯源过程中精确数据收集及产生的海量信息存储、数据类型识别、数据信息匹配等内容的研究。如何将数据溯源技术更好地融合进更多的领域以更有效地利用数据是未来需要进一步研究的内容。

表1 各领域应用对比情况

Table 1 Applications in Various Fields

应用特点 / 应用领域	作用	使用技术/模型
重大突发事件	舆情管控	区块链，人工智能，大数据等
电子商务	商品溯源，防信息篡改	区块链等
企业经营	数据管理，指标管理	W7，OPM等
科学研究	数据存储，数据共享	区块链，数据标识技术等

5 研究展望

当前，国内外的数据要素市场尚未完全成熟，因此数据溯源研究对于数据要素市场的培育具有重要的理论与现实意义。通过梳理国内外数据溯源相关文献可以发现：国外对数据溯源研究起步相对较早、研究成果比较丰富、研究体系业已形成、理论研究比较成熟，国内对数据溯源研究起步相对较晚、研究成果不够丰富、研究体系尚未形成、研究重点偏向实证。同时对数据的可靠性和高质量要求加快了区块链技术的发展，但现阶段对基于区块链技术开展的数据溯源还处于探索阶段，较多集中在概念层面，少量集中在特定领域和应用行业。区块链涉及专业知识的多重性、技术领域的复杂性，对如何使用区块链技术合理存储原始数据及标注信息、如何描述溯源记录并确保可信的研究内容较少。国内发展要借鉴国外研究，站在现有研究的基础上自成体系，培育出中国社会主义特色的数据要素市场。

未来研究工作，可以从以下几个方面继续开展

研究：

（1）与数据要素市场相结合，积极推进数据交付使用常态化。数据要素市场能实现数据流动的价值或使数据在流动中产生价值。然而，在数据要素市场中，数据的可复制性加大数据交易供给方的市场风险。采用区块链技术进行数据溯源将大大降低相关风险，因此将数据溯源与数据要素市场相结合，能够基于数据要素市场培育初期，培养市场数据溯源习惯，不仅有利于数据交付使用逐步常态化，更能促进数据要素市场体系的成熟。

（2）加快推进数据溯源标准工作，积极推进数据使用工作制度化。多个数据溯源模型都有相对独立、相对固定的应用情景，数据在多个溯源模型中如何转换缺乏规范标准。只有加快建立和完善数据溯源标准规范，不同溯源系统之间才会形成标准接口，才能实现不同类型的数据交换，才能从不同数据溯源模型中快速获取和精准收集有效的数据信息。因此，建立和完善数据溯源标准规范，要充分考虑数据溯源信息记录、描述的实用性和操作性，要考虑溯源模型后续的可扩展性、可适用性，更要加强溯源系统访问接口、数据转换技术的标准化研究。

（3）不断提升数据溯源信息质量，积极推进数据服务优质化。推进数据溯源标准规范应用力度，规范溯源信息全面采集、过程管理和提交时限，确保数据溯源信息记录的完整性、时效性和规范性。创新面向特定领域、应用情景的数据溯源验证方法和修复方法，基于数据信息的可关联性、可溯源性对溯源过程中的碎片信息进行修复，支持碎片信息的可信度评价。

（4）高度重视数据溯源信息安全，积极推进数据信息使用规范化。推进数据溯源信息记录、采集、存储、共享等全过程全生命周期标准化安全管理，以安全管理规范化、标准化降低数据溯源信息泄露风险。强化高等学校、科研院所与高新技术企业之间的产学研合作，联合申报国家科技部、工信部重大课题，对数据溯源共性关键技术在数据确权及数据交易中的安全应用、公众隐私数据保护策略进行协同攻关。

（5）高标准搭建数据溯源平台，积极推动数据要素市场健康化发展。当前，基于数据交易创新溯源技术和搭建溯源平台的研究较为缺乏。采用区块链、智能合约、人工智能等安全算法，将数据溯源嵌入数据采集、数据确权、数据流通、数据交易、数据监管等全生命周期节点，高起点谋划、高标准建设互联、互通、开放、共享的数据溯源平台。明确数据溯源平台市场定位，制定数据溯源平台运行机制，将会有力保障数据要素安全流动，在提高安全监管效率的同时促进数据交易有序进行。

参考文献：

[1] Foster I, Vockler J, Wilde M, et al. Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation[C]//Proceedings of the 14th International Conference on Scientific and Statistical Database Management. IEEE, 2002: 37-46.

[2] 如何看待数据模型在数据管理中的位置？[EB/OL]. [2019-11-02]. https://zhuanlan. zhihu. com/p/75883955. (How to View the Position of Data Model in Data Management?[EB/OL]. [2019-11-02].https://zhuanlan.zhihu.com/p/75883955.)

[3] Buneman P, Khanna S, Wang-Chiew T. Why and Where: A Characterization of Data Provenance[A]//Database Theory — ICDT[M]. Springer Berlin Heidelberg, 2001:316-330.

[4] Green T J, Karvounarakis G, Tannen V. Provenance[C]// Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2007: 31-40.

[5] Ram S, Liu J. A New Perspective on Semantics of Data Provenance[C]//Proceedings of the 1st International Conference on Semantic Web in Provenance Management - Volume 526. 2009: 35-40.

[6] 王逢阳, 徐全军, 刘峰, 等. 科学数据溯源描述模型及规范设计与思考[J]. 科研信息化技术与应用, 2017, 8(1): 27-34.(Wang Fengyang, Xu Quanjun, Liu Feng, et al. Design and Thinking of Scientific Data Provenance Description Model and Specification [J]. e-Science Technology & Application, 2017, 8(1): 27-34.)

[7] 沈志宏, 张晓林. 语义网环境下数据溯源表达模型研究综述[J]. 现代图书情报技术, 2011(4): 1-8.(Shen Zhihong, Zhang Xiaolin. Data Provenance Model in Semantic Web Environment: An Overview [J]. New Technology of Library and Information Service, 2011(4): 1-8.)

[8] Provenance Vocabulary Mappings[EB/OL]. [2012-06-30]. http://www. w3. org/2005/Incubator/prov/wiki/Provenance_Vocabulary Mappings.

[9] Groth P, Moreau L. PROV-Overview. An Overview of the PROV Family of Documents[R]. Southampton, UK: W3C, 2013.

[10] 倪静, 孟宪学. PROV数据溯源模型及Web应用[J]. 图书情报工作, 2014, 58(3): 13-19.(Ni Jing, Meng Xianxue. PROV Model

and Its Web Application[J]. Library and Information Service, 2014, 58(3): 13-19.)

[11] 倪静, 孟宪学. 关联数据环境下数据溯源描述语言的比较研究[J]. 现代图书情报技术, 2013(2): 18-23.(Ni Jing, Meng Xianxue. The Comparative Analysis of Major Provenance Vocabularies in Linked Data Environment[J]. New Technology of Library and Information Service, 2013(2): 18-23.)

[12] GB/T 34945-2017 信息技术 数据溯源描述模型[EB/OL]. https://max. book118. com/html/2018/1203/7054141150001162. shtm. (GB/T 34945-2017 Information Technology Data Traceability Description Model[EB/OL]. https://max.book118.com/html/2018/1203/7054141150001162.shtm.)

[13] Sahoo S S, Barga R S, Goldstein J, et al. Provenance Algebra and Materialized View-based Provenance Management[C]//Proceedings of the 2nd International Provenance and Annotation Workshop. Berlin: Springer, 2008: 531-540.

[14] 杜莹, 林冰仙, 周良辰, 等. 面向SAR数据处理流程的溯源方法研究[J]. 武汉大学学报·信息科学版, 2017, 42(5): 669-675.(Du Ying, Lin Bingxian, Zhou Liangchen, et al. Provenance Method for SAR Data Processing Flow[J]. Geomatics and Information Science of Wuhan University, 2017, 42(5): 669-675.)

[15] 袁洁. 基于关联数据技术的空间数据溯源共享研究[D]. 武汉: 武汉大学, 2013. (Yuan Jie. Research on Geospatial Data Provenance Sharing Based on Linked Data Approach[D]. Wuhan: Wuhan University, 2013.)

[16] Hasan R, Sion R, Winslett M. Introducing Secure Provenance: Problems and Challenges[C]//Proceedings of the 2007 ACM Workshop on Storage Security and Survivability. New York: ACM Press, 2007: 13-18.

[17] 李秀美, 王凤英. 数据起源安全模型研究[J]. 山东理工大学学报(自然科学版), 2010, 24(4): 56-60. (Li Xiumei, Wang Fengying. Research on Data Provenance’s Security Model[J]. Journal of Shandong University of Technology(Natural Science Edition), 2010, 24(4): 56-60.)

[18] 王凤英, 张方, 张伟. 基于医疗健康大数据的安全起源模型与可信性验证算法[J]. 山东理工大学学报(自然科学版), 2017, 31(6): 6-11. (Wang Fengying, Zhang Fang, Zhang Wei. Securing Data Provenance and Creditability Validation Study Based on Big Data of Health Care[J]. Journal of Shandong University of Technology (Natural Science Edition), 2017, 31(6): 6-11.)

[19] 邓仲华, 容益芳. 一种分层次的数据溯源安全模型[J]. 图书馆学研究, 2016(20): 36-41. (Deng Zhonghua, Rong Yifang. A Hierarchical Data Traceability Security Model[J]. Researches in Library Science, 2016(20): 36-41.)

[20] 刘耀宗, 刘云恒. 基于区块链的RFID大数据安全溯源模型[J]. 计算机科学, 2018, 45(S2): 367-368, 381. (Liu Yaozong, Liu Yunheng. Security Provenance Model for RFID Big Data Based on Blockchain[J]. Computer Science, 2018, 45(S2): 367-368,381.)

[21] Liang X P, Shetty S, Tosh D, et al. ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability[C]//Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing(CCGRID). IEEE, 2017: 468-477.

[22] 王芳, 赵洪, 马嘉悦, 等. 数据科学视角下数据溯源研究与实践进展[J]. 中国图书馆学报, 2019, 45(5): 79-100. (Wang Fang, Zhao Hong, Ma Jiayue, et al. Research and Practice Progress of Data Provenance from the Perspective of Data Science[J]. Journal of Library Science in China, 2019, 45(5): 79-100.)

[23] 周忠. 数据起源技术研究及其在PostgreSQL中的实现[D]. 广州: 华南理工大学, 2016. (Zhou Zhong. A Research of Data Provenance Technology and Its Implementation in PostgreSQL [D]. Guangzhou: South China University of Technology, 2016.)

[24] Karvounarakis G, Ives Z G, Tannen V. Querying Data Provenance [C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. 2010: 951-962.

[25] 王黎维, 鲍芝峰, Koehler Henning, 等. 一种优化关系型溯源信息存储的新方法[J]. 计算机学报, 2011, 34(10): 1863-1875. (Wang Liwei, Bao Zhifeng, Koehler Henning, et al. An Approach for Optimizing Relational Provenance Storage[J]. Chinese Journal of Computers, 2011, 34(10): 1863-1875.)

[26] Deutch D, Milo T, Roy S, et al. Circuits for Datalog Provenance [C]//Proceedings of International Conference on Database Theory. 2014: 201-212.

[27] Chapman A P, Jagadish H V, Ramanan P. Efficient Provenance Storage[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008: 993-1006.

[28] 王黎维, 黄泽谦, 罗敏, 等. 集成对象代理数据库的科学工作流服务框架中的数据跟踪[J]. 计算机学报, 2008, 31(5): 721-732. (Wang Liwei, Huang Zeqian, Luo Min, et al. Data Provenance in a Scientific Workflow Service Framework Integrated with Object Deputy Database[J]. Chinese Journal of Computers, 2008, 31(5): 721-732.)

[29] 吴渊. 工作流系统-Nebulas中数据溯源框架的设计与实现[D]. 昆明: 昆明理工大学, 2011. (Wu Yuan. Design and Implementation of a Provenance Framework in Workflow System-Nebulas[D]. Kunming: Kunming University of Science and Technology, 2011.)

[30] 邓仲华, 魏银珍. 面向数据发布的科学工作流数据溯源方法研究[J]. 图书与情报, 2014(3): 61-66. (Deng Zhonghua, Wei Yinzhen. Study on the Method of Provenance in Science Workflow for Data Publishing[J]. Library & Information, 2014 (3): 61-66.)

[31] Billings J J. Applying Distributed Ledgers to Manage Workflow Provenance[OL]. arXiv Preprint, arXiv:1804.05395.

[32] 魏银珍, 邓仲华. 云环境下科学工作流的溯源数据收集和查询框架研究[J]. 情报理论与实践, 2015, 38(7): 115-118. (Wei Yinzhen, Deng Zhonghua. Research on Data Provenance

Collection and Query Framework of Scientific Workflow in Cloud Environment[J]. Information Studies: Theory & Application, 2015, 38(7): 115-118.)

[33] Park H, Ikeda R, Widom J. RAMP: A System for Capturing and Tracing Provenance in MapReduce Workflows[C]//Proceedings of the 37th International Conference on Very Large Data Bases (VLDB 2011). 2011: 1351-1354.

[34] Saad M I M, Jalil K A, Manaf M. Data Provenance Trusted Model in Cloud Computing[C]//Proceedings of 2013 International Conference on Research and Innovation in Information Systems (ICRIIS). IEEE, 2013: 257-262.

[35] Zawoad S, Hasan R. SECAP: Towards Securing Application Provenance in the Cloud[C]//Proceedings of IEEE 9th International Conference on Cloud Computing. IEEE, 2016: 900-903.

[36] Tosh D K, Shetty S, Liang X P, et al. Consensus Protocols for Blockchain-Based Data Provenance: Challenges and Opportunities [C]//Proceedings of the 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference. IEEE, 2017: 469-474.

[37] Kim H M, Laskowski M. Towards an Ontology-Driven Blockchain Design for Supply Chain Provenance [OL]. arXiv Preprint, arXiv:1610.02922.

[38] 王若佳, 李培. 基于互联网搜索数据的流感监测模型比较与优化[J]. 图书情报工作, 2016, 60(18): 122-132.(Wang Ruojia, Li Pei. Detecting Influenza Epidemics by Comparing and Optimizing Models Based on Internet Search Engine Query Data [J]. Library and Information Service, 2016, 60(18): 122-132.)

[39] 王迪, 杨广义. 基于区块链溯源技术重大公共卫生事件的舆情防控研究[J]. 河北工程大学学报(社会科学版), 2021, 38(1): 30-33.(Wang Di, Yang Guangyi. Research on Public Opinion Risk Control of Major Public Health Problems Based on Blockchain Traceability Technology[J]. Journal of Hebei University of Engineering (Social Science Edition), 2021, 38(1): 30-33.)

[40] 人工智能助力疾病预测 平安科技携手重庆疾控联合研发全球首创人工智能+大数据流感预测模型[EB/OL]. [2017-07-25]. http://www. pingan. cn/zh/common/cn_news/1500961992328. shtml. (AI Assisted Disease Prediction PingAn Science and Technology Cooperates with Chongqing CDC to Jointly Develop the First Global AI + Big Data Influenza Prediction Model[EB/OL]. [2017-07-25]. http://www. pingan. cn/zh/common/cn_news/1500961992328.shtml.)

[41] 朱鹏, 朱星圳, 王莉, 等. 基于时间序列与信息融合的突发事件信息瀑布溯源方法[J]. 现代情报, 2018, 38(10): 38-42.(Zhu Peng, Zhu Xingzhen, Wang Li, et al. Tracing Method of Emergencies Information Cascade Based on Time Series and Information Fusion[J]. Journal of Modern Information, 2018, 38 (10): 38-42.)

[42] 陈卫哨. 微博突发事件检测及溯源技术研究[D]. 哈尔滨: 哈尔滨工程大学, 2014.(Chen Weishao. Burst Event Detection and Initialyzing Technology Research in Micro-Blog[D]. Harbin: Harbin Engineering University, 2014.)

[43] 京东万象以科技助力数据流通, 采用区块链技术促行业健康发展 [EB/OL]. [2017-01-11]. https://wx. jdcloud. com/resources/preview/58? winzoom=1. (Jingdong Vientiane Uses Science and Technology to Facilitate Data Circulation and Uses Blockchain Technology to Promote Healthy Development of the Industry[EB/OL]. [2017-01-11]. https://wx.jdcloud.com/resources/preview/58?winzoom=1.)

[44] 缪新萍, 吴漾, 孔庆波, 等. 电网企业指标数据溯源模型研究与设计[J]. 电力大数据, 2021, 24(4): 70-77.(Miao Xinping, Wu Yang, Kong Qingbo, et al. Research and Design of Index Data Provenance Model for Power Grid Enterprises[J]. Power Systems and Big Data, 2021, 24(4): 70-77.)

[45] 王姝, 孙善鹏, 樊景超, 等. 基于区块链的农业科学数据溯源应用初探[J]. 农业大数据学报, 2020, 2(2): 47-54.(Wang Shu, Sun Shanpeng, Fan Jingchao, et al. Preliminary Study on the Traceability Application of Agricultural Science Data Based on Blockchain[J]. Journal of Agricultural Big Data, 2020, 2(2): 47-54.)

[46] 英国数字保存中心发布指南《制定数据管理与共享计划》[EB/OL]. [2011-11-17]. http://www. ecas. cas. cn/xxkw/kbcd/201115_83713/ml/xxhzlyzc/201111/t20111117_3397761. html. (GuideLines Issued by the British Digital Preservation Centre 'Developing Data Management and Sharing Plan'[EB/OL].[2011-11-17].http://www.ecas.cas.cn/xxkw/kbcd/201115_83713/ml/xxhzlyzc/201111/t20111117_3397761.html.)

[47] 谷俊, 许鑫. 人文社科数据共享模型的设计与实现: 以联盟链技术为例[J]. 情报学报, 2019, 38(4): 354-367.(Gu Jun, Xu Xin. Design and Implementation of a Humanities and Social Sciences Data Sharing Model: A Case Study of Consortium Blockchain[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(4): 354-367.)

作者贡献声明：

王晓庆：提出研究思路，设计研究方案；
孙战伟：起草论文；
吴军红：信息检索，信息分析；
杜自然：选题设计，内容修改；
钱城江：论文最终版本修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

收稿日期：2021-12-15
收修改稿日期：2022-01-10

Research Progress of Data Traceability from the Perspective of Data Element Circulation

Wang Xiaoqing[1, 2, 3] Sun Zhanwei[1] Wu Junhong[4] Du Ziran[5] Qian Chengjiang[6]

[1](School of Public Administration, Nanjing University of Finance & Economics, Nanjing 210003, China)

[2](College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

[3](Hongshan College, Nanjing University of Finance & Economics, Nanjing 210003, China)

[4](Department of Platform Research and Development, Business School, Nanjing Normal University, Nanjing 210023, China)

[5](Department of Platform Research and Development, Greater Bay Area Big Data Research Institute, Shenzhen 518048, China)

[6](Nanjing NJtech Safety Co., Ltd, Nanjing 210047, China)

Abstract: [**Objective**] The research progress and application scenarios of data traceability are analyzed through literature review, in order to provide reference for the construction of data trading platform, the construction of industrial data governance and the construction of digital government governance. [**Methods**] The data traceability model, data traceability method and data traceability application are summarized and analyzed, and on this basis, the research status and shortcomings are discussed. [**Results**] Whether in content description, model construction or scene application, data traceability research has achieved rich results, such as improving the quality of data traceability, ensuring the safety of data traceability and improving the efficiency of data traceability. [**Limitations**] The research on data traceability from the perspective of factor circulation started relatively late, the research results were not rich enough, the research system had not been formed, and the research focus was biased towards empirical research. [**Conclusions**] We can actively promote the normalization of data delivery and use by combining with data factor market; speed up the work of data traceability standards, and actively promote the institutionalization of data use; continuously improve the quality of data traceability information, and actively promote the quality of data services; attach great importance to data traceability information security, and actively promote the standardization of data information use; to build a high standard data traceability platform, and actively promote the healthy development of data factor market.

Keywords: Data Circulation Data Traceability Management Model Data Factor