DOI:10.13530/j.cnki.jlis.190043

# 数据科学视角下数据溯源研究与实践进展*

王 芳 赵 洪 马嘉悦 李晓阳 张晓玥

**摘 要** 真实性和可靠性是当前各领域对数据的根本要求,基于数据溯源实现数据的质量控制与可信管理具有重要的研究价值和实践意义。数据溯源不仅是一个技术问题,同时也是一个管理问题,在数据科学范式下应当受到信息资源管理研究的关注和重视。鉴于此,本文结合相关领域的最新研究进展,系统阐述了数据溯源的概念发展与内涵;梳理了面向数据溯源管理的信息描述模型、通用表达模型、领域应用模型、安全管理模型与区块链溯源管理模型;描述了关系数据库、科学工作流、大数据平台、云计算和区块链等典型应用环境下的数据溯源计算方法。此外,本文还重点分析了数据溯源在数字图书馆、档案信息管理、网络信息资源管理、科学数据共享管理及电子商务信息系统等信息资源管理研究领域中的应用价值与相关实践,并对数据溯源技术方法、标准规范、信息安全、区块链融合以及模型扩展验证等方面的发展进行了展望,以期为数据管理和数据科学领域的研究人员提供参考。参考文献 136。

**关键词** 数据科学 数据溯源 区块链 数据质量 大数据平台

**分类号** G203

# Research and Practice Progress of Data Provenance from the Perspective of Data Science

WANG Fang, ZHAO Hong, MA Jiayue, LI Xiaoyang & ZHANG Xiaoyue

**ABSTRACT**

In the data age, authenticity and reliability are the fundamental requirements of data in many fields. It is of great research value and practical significance to realize data quality control and reliable management through data provenance. Data provenance is not only a technical problem but also a management problem. It should be paid more attention to by scholars in the field of data science and information resources management.

Data provenance is widely applied in scientific data curation, e-commerce, food safety, culture and art, medical treatment, digital library, electronic document management and many other fields, and a lot of studies on it have been conducted. From the perspective of data science, this paper reviews the research and practice progress of data provenance based on 136 domestic and foreign research papers. On the basis

* 本文系国家社会科学基金重大项目“我国网络社会治理研究”(编号:14ZDA063)与提升政府治理能力大数据应用技术国家工程实验室开放基金重点支持项目“基于 NLP 和深度学习的大规模政府公文智能处理技术研究”的研究成果之一。(This article is an outcome of the major project “Governance of China Network Society” (No. 14ZDA063) supported by National Social Science Foundation of China and the project “Intelligent Processing Technology for Large-scale Government Documents Based on NLP and Deep Learning” supported by CETC Big Data Research Institute Co., Ltd.)

通信作者:王芳,Email:wangfangnk@nankai.edu.cn,ORCID:0000-0002-2655-9975 (Correspondence should be addressed to Wang Fang, Email: wangfangnk@nankai.edu.cn, ORCID: 0000-0002-2655-9975)

of reviewing the concepts, models, computation methods and practical applications of data provenance, this paper introduces the related studies and practice in the field of information resources management. Finally, the future research trends on data provenance are discussed.

This paper systematically combs the development of the concept of data provenance and introduces five types of data provenance models according to their function level and application characteristics in data management, including information description model, general expression model, domain application model, safety management model and block-chain provenance management model. Model is the abstract representation framework of the strategy and process of data provenance, while computation is the technique and algorithm of its implementation. The computation of data provenance can be divided into two basic ideas: tag-based provenance and non-tag-based provenance. Some specific computing methods have been developed for different application scenarios. This paper mainly introduces the computing methods in typical application scenarios such as relational database, scientific workflow, big data platform, cloud computing and block-chain. This paper also focuses on the research and practice of data provenance in the fields of digital library, archival information management, online information resources management, scientific data sharing and curation as well as electronic commerce information system.

On the whole, there are still some limitations in the research on data provenance technology, standards and specifications, information security, block-chain fusion and model extension and verification. In the future, more in-depth research and practical exploration are needed in these areas. This paper is expected to help scholars of information resources management and data science have a comprehensive understanding of data provenance. The limitation of this paper is that it fails to formulate a data provenance model for information resources management within the required length, which will be conducted independently in future research. 136 refs.

**KEY WORDS**

Data science. Data provenance. Blockchain. Data quality. Big data platform.

## 0 引言

在大数据环境中,数据的生成规模、传播速度急剧增长,数据的来源以及衍生路径呈现出多样化、复杂化的特点,给传统静态的信息资源管理方式带来挑战。众多领域,如科学研究、应急管理、决策支持、信用监管、循证医学、犯罪证据等,对于数据的真实性与有效性有着很高的要求。但是,“许多数据不是直接测量获得的,而是通过计算机程序的应用从其他数据中得到的”[1]。在数据的共享、交换与交易过程中,原生数据常常经过多次流转、复制、迁移、集成、抽取、计算等操作后形成海量派生数据,如果不对原生数据的溯源信息进行记录,将在很大程度上降低派生数据的真实性和有效性[2],以致给数据的应用带来风险。为了增加数据的可靠性和可信性,追溯数据的派生过程和原生系统,数据溯源技术应运而生。

数据溯源最早应用于 e-Science 领域,具体涉及地理、生物、天文、历史、考古等对数据真实性要求较高的学科[3]。当对科研结果产生质疑时,常常需要对作者的研究数据进行回溯,以发现错误源并对研究结果进行修正[4]。在 e-Science 中,数据管护(Data Curation)通过全生命周期管理来提高科学数据的可复用性与价值,溯源元数据就是其中一个重要手段[5]。在食品安全领域,食品生产过程信息的记录可以保证食品

安全[6];在文化艺术领域,需要借助工艺品的先期所有权和溯源信息鉴定其真实性[7];在医疗领域,溯源信息对于 DNA 检测、药物试验和医疗责任确定等至关重要[8]。随着信息技术的发展,数据溯源又扩展到数据库、工作流、语义网、普适计算、网格计算和无线传感器等计算机研究领域。

随着诸多领域对数据溯源的应用需求日益迫切,相关研究也越来越深入。自 2005 年以来,一些学者对数据溯源相关的研究成果进行了梳理。Simmhan 等对数据溯源技术进行了分类,并比较各类溯源技术在 e-Science 项目中的应用[9];刘喜平等详细梳理了数据溯源的查询反演方式和标注方式的特点及其应用[10];戴超凡等对已有成熟的溯源管理系统进行了分析和比较[11];高明等以模式级和实例级数据溯源的表示、查询为主线综述了数据溯源的研究进展[12];沈志宏等综述了语义网环境下数据溯源在表达模型与技术上的研究进展[2];徐杨等概述了数据溯源的模型、方法、技术和应用系统状况[13];明华等总结了数据溯源的 7 种技术模型[3];Herschel 等概括了数据溯源的类型和生成方法,重点关注了溯源的用途、应用形式和方案选择[14];石丽波等详述了数据溯源安全研究,重点介绍了起源过滤和起源感知的访问控制机制的主要特征[15]。这些文献大部分发表于 2011 年之前,对了解数据溯源的主要内容和发展历史具有很好的参考价值。但随着数据溯源实践应用的进一步发展,特别是云计算、大数据以及区块链技术的深入应用,数据溯源的研究与应用取得了新的进展,成为当前数据科学发展中不可忽视的组成部分,亟需进行全面梳理与系统综述。

数据是信息资源管理的重要对象,数据的真实性、可靠性与完整性更是档案学电子文件管理的首要标准,数据管护、开放存取与机构知识库等研究的深入使图书情报学对数据质量的要求日益增长。为此,本文拟结合最新研究进展,从数据科学视角系统梳理数据溯源的概念内涵、模型及方法,分析数据溯源在信息资源管理中的应用价值与相关实践,并对未来的研究方向进行展望。

## 1 数据溯源概念内涵

数据溯源(Data Provenance)的概念经历了一个不断发展完善的过程。1990 年,Wang 等提出异构数据库系统中“数据来自哪里”和“哪些中间数据来源被用来实现数据”的问题,是产生数据溯源思想的萌芽[16]。1991 年,Lanter 首次提出“Data Lineage”(数据世系)一词,用以对目标数据的源头数据和应用转化过程进行描述[17]。2000 年,Cui 等从关系操作角度出发,定义了视图 Data Lineage,希望能够标识数据仓库视图中给定数据项的源头数据项集[18]。2000 年,Buneman 等在研究数据溯源时使用了“Data Provenance”(数据溯源)一词[19];2001 年他又将数据溯源术语范围扩大,指出“Data Provenance”也被称为“Data Lineage”或“Data Pedigree”(数据谱系),同时明确数据溯源包括 why-provenance 和 where-provenance 两层含义,分别指对数据存在某种影响的源头数据和提取数据在源头数据库中的位置[20]。至此,数据溯源(Data Provenance)开始作为正式术语使用,而且含义更加明确。

数据溯源在不同领域发展应用时,其含义也有所差别。Goble 从生物信息学角度,指出数据溯源除了要追溯目标数据的源头数据和过程信息外,还需记录数据演变过程中涉及的工具信息,需要更加详细的信息记录以保证数据的可重用性[21]。Greenwood 等认为数据溯源是一种记录信息的派生路径和注释的元数据[22]。Simmhan 等也认为数据溯源是一种元数据,用于记录数据产品最原始数据和演变转换过程的信息[9]。Glavic 等认为数据溯源包括两方面,一方面是将数据项的溯源描述为导致其创建的过程,另一方面则关注派生数据项的原始数据来源[23]。Braun 等认为数据溯源通过标注和连接描述对象产生当前状态的过程,从而表示出各种参与对象间的因果关系[24]。戴超凡等将数据溯源定义为

记录数据从产生到消亡或转换的整个生命周期内所发生的变化和经过处理的信息[11]。

上述定义展示了数据溯源内容的不同侧重点,但均关注源头数据和过程转化信息。很多定义将数据溯源定义为一种特殊类型的元数据,用于描述数据的源头、创建和流转过程等环节。但我们认为,描述性元数据只能构成数据溯源的一部分内容,并不能将数据溯源的整个过程表达出来,数据溯源概念需要在多层面上进行多视角的理解。从"数据"层面上讲,数据的形式是多种多样的,在关系数据库中,溯源的数据可以是视图、表、元组、属性或数据库中更细粒度的数据项,甚至还可以是指向外部的数据资源[9];在跨系统的异构数据中,溯源的数据可以是文本文件、图片、音频、视频和 XML 等多种形式。从"溯源(Provenance)"层面上讲,溯源是描述参与生产和交付数据或事物中涉及的实体、活动和人员的信息[24];可以描述最终产品生产过程中从数字数据到物理对象的任何信息[15];也可以将它看作是一种元数据,它跟踪数据派生的步骤,在数据密集型场景中可以提供重要的附加价值[9]。

因此,本文认为数据溯源是对目标数据的源头数据以及流转过程中的变动加以追溯、确认、描述和记录保存的过程。主要包括三方面内容:一是对产生当前数据项的源头数据的追溯与描述;二是对源头数据如何演变为当前数据状态的过程信息的追溯、捕获或记录,包括数据的移动、变化、交付、执行、操作、传播和利用等行为动作,以及在这一过程中产生的派生数据内容;三是在从源头数据到当前数据状态这一过程中,对所有能够影响数据状态的因素(比如影响数据的实体、工具等)进行追溯、描述和记录。因而,数据溯源既可能是一种从数据当前状态向源头数据追溯的逆向运动过程,也可能是从源头数据运动之始就对数据变动信息进行捕获和记录的过程。追溯过程中所记录下来的信息,形成了一套内容丰富、全面系统、联系紧密的数据集,也就是数据溯源的结果信息。

## 2 数据溯源管理模型

选择或建立溯源信息的管理模型是数据溯源的基础,用以确定数据溯源的基本思路、基础框架和主要步骤。目前,不同行业、不同领域、不同组织中产生了许多数据溯源管理的模型,这些模型基于不同的任务目标构建,有些具有较好的通用性,有些则用于特定的数据对象或应用场景。本文根据溯源模型在数据管理中的作用层次与应用特点,将其划分为数据溯源信息描述模型、数据溯源通用表达模型、数据溯源领域应用模型、数据溯源安全管理模型与区块链溯源管理模型等五类。

### 2.1 数据溯源信息描述模型

数据溯源信息描述模型是为了获取和记录溯源信息而构建的,溯源信息的内容描述是数据溯源管理的基础。最早的研究关注数据事件来自哪里(where),一般的做法是将来源注释信息作为数据的扩展元数据进行共同存储[16],但面对复杂数据时易产生管理混乱问题。由此,Buneman 等将溯源信息进行专门的分类描述,将其划分为 where 类元素和 why 类元素,增加了数据事件发生的原因描述(why)[20],并进一步扩充了 how 类元素,用以记录数据事件发生的动作[25]。这种包含<where, why, how>三类元素的 W3 模型,在数据库领域中应用较为广泛,较大程度上满足了基于数据库的数据溯源需求。

与基于数据库的数据溯源相比,基于工作流的数据溯源对溯源信息描述提出了更高的要求,其溯源信息旨在帮助参与者验证数据、重现过程、优化流程、评估质量和挖掘关系等。针对 W3 模型不能完全适用于工作流领域的问题,Ram 等提出了 W7 模型,将溯源信息的内容描述为七类相互关联的元素:<what, when, where, how, who, which, why>,每个元素都可以追踪数据生命周期中影响数据的因素,如事件(event)、时间(time)、地点(location)、行动(action)、代理

(agent)、设备(device)与原因(reason),达到工作流数据溯源的目的[26]。

W7 模型虽然可以溯源工作流中的绝大部分关键信息内容,但在科学数据流中表示科研过程信息时还存在不精细的问题。为此,王逢阳等将科学数据的特性与 W7 模型相结合,提出了"W7+R3"科学数据溯源内容模型[27],其中 R3 为 Reference、Result 和 Remark。Reference 记录事件发生的参考资料信息,Result 记录事件发生的实际结果及其成因分析,Remark 描述事件发生的其他重要备注信息。该模型在科研准备阶段、科研实施阶段和科研结果中,各元素分别记录数据处理事件的相应信息。

但是,并非所有环境下的数据溯源都需要完整采用 W7 模型。一方面,W7 模型在详细记录溯源信息的同时,也带来了很大的存储开销;另一方面,在关联数据和语义网等环境下,数据集的元数据溯源信息本身也是元数据,如在 DC 元数据基础上扩展的"DCMI 元数据术语"就包括描述性元数据与溯源性元数据共计 55 个属性,以避免多次重复的发布过程。此外,基于本体的模型以自定义类和属性的方式,为不同应用和不同领域的溯源信息建模,并形成特定的溯源信息交换与管理框架,因而通常也只是支持 W7 模型的一部分[28]。

### 2.2 数据溯源通用表达模型

数据溯源通用表达模型的目标是通过建立一套完整的技术规范,使数据溯源管理全过程遵循通用的表达方式和共性的描述标准,并提供模型的扩展方法,从而适用于各类溯源管理应用场景。第一个典型的通用表达模型是首届国际溯源和标注组织(International Provenance and Annotation Workshop, IPAW)会议颁布的 OPM 模型(Open Provenance Model,开放溯源模型),已成为数据溯源信息交换的标准[29]。OPM 模型定义了一组核心规则,用来描述事务间关系,包括 Artifact(实体)、Process (过程) 和 Agent (代理) 3 类节点。这三类节点中存在五种关系,分别是 used (使用)、was Generated By (生成)、was Controlled By(控制)、 was Triggered By(触发)、was Derived From(获得),一般通过构建溯源模型图来描述节点之间的关系[30]。

OPM 模型溯源标准在不断推广的同时,也存在一些问题,如概念术语和用法的模糊性或概念本身设计不恰当等。因而,W3C (World Wide Web Consortium,万维网联盟)在对 OPM 进行重大修改后,于 2013 年 4 月 30 日发布了 PROV 溯源标准,并出台了面向用户、开发人员和高级开发人员的一系列数据溯源相关的文档[31]。它的核心是概念数据模型 PROV-DM,该模型的结构包括 Entity(实体)、Activity(活动)、Agent(代理)三类以及七种类间关系集[32]。PROV 模型可以描述影响实体不同状态的所有活动及其关系链,由此可以追溯到每个实体在不同阶段的溯源信息。然而,标准 PROV 模型在对数据语义关系加强描述的同时,也使得构造的数据模型过于复杂,每一个实体间关系都关联了模型中全部定义的关系类型。因此,在具体应用过程中,还需要对数据模型进行优化,在完整保存溯源信息的同时减少数据表数量。

在国内,GB/T 34945-2017《信息技术 数据溯源描述模型》国家标准于 2017 年 11 月正式发布,定义了 ProVOC (Provenance Vocabulary Model)数据溯源模型[33]。该模型由数据、活动和执行实体 3 个一级类构件组成。"数据"包括"参数"和"数据集"两个二级子类构件,"数据集"按应用领域可分解出若干三级或更多层级的子类(特定数据集);"活动"是由执行实体发起或受执行实体控制、影响的具体动作,包括人类执行实体(如:人、组织和公司)和非人类执行实体(如:数据提供之类的服务)等二级子类构件。目前的 ProVOC 模型还只是一个轻量级溯源模型,所制定的标准主要为各类构件描述信息。尽管该模型可以支持在特定领域上的灵活扩展,但现阶段的应用范围仍然较为狭小,通用性和管理效果也有待进一步验证和评价,未来需在标准的实施推广过程中不断完善。

### 2.3 数据溯源领域应用模型

在数据溯源表达的通用标准不断完善的同时,许多领域根据其特定的应用场景和溯源目标,提出和发展了许多特定的数据溯源模型,对通用表达模型进行了有效补充、扩展和深化。如第二届 IPAW 会议发布的基于本体的溯源信息管理模型——Provenir 模型[34],通过自定义类间关系构成模型框架及逻辑描述,在生命科学、海洋、传感器和卫生保健等领域都得到了广泛应用[13]。

在特定应用领域,Wang 等开发了面向医疗系统的 TVC (Time-and-Value Centric)溯源模型[35],它将事务性系统中基于时间和基于标注溯源的优点相结合,是一个简单高效的混合型溯源模型。陈颖借鉴了生物信息学中的 DNA 双螺旋结构来构建数据溯源模型,支持科学工作流的数据动态重用及再执行,直观揭示数据序列的变化[36]。Kulkarni 为关键值数据库设计了一种关键值溯源模型 KVPM (Key-Value Provenance Model)[37],根据键值系统中不同密钥对应的不同属性,可以随时添加或删除任何键的属性,灵活选择数据模型层次结构,在不同层次上收集溯源信息。杜莹等提出了基于分层二部图的溯源模型 HBPM (Hierarchical Bigraph Provenance Model)[38],该模型与溯源信息的收集及处理机制相吻合,适合以复杂高密度数据处理流程为主且处理参数复杂的应用场景。此外,针对空间数据溯源的需求,袁洁基于数据关联技术设计空间数据溯源模型,并引入本体实现空间数据溯源模型的形式化表达[39]。

除了特定领域的应用模型外,面向领域建模的数据溯源模型也在不断发展。在第三届 IPAW 会议中,Hartig 提出的关联数据环境下的 PV (Provenance Vocabulary)模型[40],将溯源信息集成到具有质量概念的关联数据中,Web 发布时同步生成溯源信息。VoID (Vocabulary of Interlinked Datasets)模型也是一种用于描述关联数据集的数据溯源模型[41],本质上也是一组描述元数据,是关联数据集溯源信息的元数据表达和存储基础[28],用于发现如 DBPedia 等海量关联数据集的可用信息。Omitola 等人进一步延伸出了 VoIDP (Vocabulary of Interlinked Datasets Provenance)模型,用于发布时添加溯源元数据信息[42]。基于众多词表,W3C 组织形成了一个概括的溯源本体 PROV-O,将术语(类和属性)定义为三类:起点术语、扩展术语和限定关系的术语,以渐增方式创建本体,不仅可以直接用于表示溯源信息,还可以作为创建特定领域溯源本体的参考模型[43]。

### 2.4 数据溯源安全管理模型

数据溯源能够对数据处理的来源和流程信息进行管理,但溯源信息本身也是一种数据,在不可信环境中经过不同的用户和处理时,数据溯源信息极易被转换、编辑甚至恶意篡改,存在很大的安全隐患。传统的安全模型主要针对外部防御,通过控制用户访问和修改数据的权限,从而保障数据的安全性。但是,这些安全措施无法防止内部授权用户由于种种原因对数据进行的各类操作处理。为了使溯源信息能够充分体现完整性、可靠性和可用性,需建立数据溯源安全管理模型,从外部和内部共同保护溯源信息的安全。

Hansan 等认为安全可信源机制能够检测数据溯源链的完整性,判断是否被明显破坏,还能确保溯源内容是保密的,而且管理员不必知道内容就可以验证其真实性[44];并基于多密钥拷贝可信审计与广播加密机制创建分析数据溯源威胁模型,以保障文件系统溯源信息的完整性和机密性[45]。但这种静态逻辑密钥分配的加密机制,存在无法动态扩展的问题,也无法将模型直接应用于数据库。Zhang 等改进的数据溯源威胁模型则解决了数据库中溯源信息完整性管理的两个问题,即对每条记录添加溯源检查,并将数据处理的线性序列管理转变为有限非循环图的管理方式[46]。李秀美进一步对数据溯源威胁模型进行补充,提出一种引入了时间戳参数和逻辑密钥分层管理方案的数据溯源安全模型,保护数据生命周期中按时间组成的溯源链的完整性和机密性,并应用于医疗工作流系

统[47]。Park 等基于依赖名(Dependency Names)和依赖列表(Dependency Lists)的概念,为每个依赖列表项定义一个溯源的依赖名,通过在访问控制策略中复用依赖名的方式简化策略规约和管理的复杂度,并以此提出了一系列基于溯源的访问控制模型 PBAC (Provenance-based Access Control),这些模型根据访问评估函数对主体和对象进行用户授权和操作验证,做出允许或拒绝访问及操作的决策,在工作流控制、原点控制、对象版本控制及动态分离的溯源任务中具有较好的安全保护效果[48]。

数据溯源安全管理还可以利用已有数据溯源标准或模型进行扩展。王凤英等基于 PROV-DM 模型构建了安全数据溯源模型 PROV-S,在 PROV-DM 模型核心结构中增加安全起源节点 S 及节点间的安全伴生关系、触发关系等关联关系[49]。邓仲华等组合 W7 模型、OPM 模型和 DNA 双螺旋结构数据溯源模型,提炼出了一种分层次的数据溯源安全模型,在模型的语义层、逻辑层和安全层三个层次中,安全层通过访问控制模式来预防无关人员对溯源信息的恶意篡改,同时也存储内部人员的访问及操作记录,保障数据对象信息的可恢复性和数据溯源信息的安全性[50]。

### 2.5 区块链溯源管理模型

区块链(Blockchain)是利用加密链式区块结构来验证与存储数据、使用分布式节点共识机制(Consensus)来生成和更新数据,并通过智能合约(Smart Contracts)来编程和操作数据的一种去中心化的基础架构与分布式计算范式[51]。区块链所具有的难伪造、防篡改、完整性和强一致性,为数据溯源管理提供了一种新的可靠解决方案。Ramachandran 等基于区块链开发了一个安全且不可变的科学数据溯源管理模型框架 SmartProvenance,该框架利用智能合约和开放溯源模型(OPM)记录溯源信息,区块链在其中作为一个平台来促进可信数据来源的收集、验证和管理,从而防止对其进行任何恶意篡改[52]。Neisse 等面向《通用数据保护条例》(GDPR)新的数据保护要求,为基于区块链的数据审计和溯源追踪方案构造了两个模型——主体合约模型(Subject Contract Model)和逻辑合约模型(Controller Contract Model),并分别评估了不同智能合约基数和数据审计粒度以及数据主体、控制器和处理器下的溯源模型实现条件、性能和扩展性[53]。刘耀宗等针对物联网中的射频识别 RFID 大数据,定义了基于区块链的物联网大数据溯源安全模型,通过引入区块链,采用去中心化的方式维护 RFID 大数据的溯源信息,实现防篡改、自带信任以及多签名复杂权限验证的分布式管理[54]。

Liang 等认为建立在区块链上的云存储数据溯源系统应具有实时云数据溯源、防篡改环境、增强的隐私保护和溯源数据验证四方面的数据操作审计能力,提出了一种基于区块链的分布式可信云数据溯源框架模型 ProvChain[55]。ProvChain 将数据嵌入到区块链交易来收集和验证云数据的溯源信息,同时保留用户身份的哈希值,以保护他们的隐私不受区块链网络中其他节点的影响。

基于区块链的数据溯源管理从存储、处理以及外部访问的可信性等方面提供了完整性和可靠性保障。尽管当前的区块链技术在数据溯源管理时所适配的复杂场景较少[56],也尚未形成基于区块链的数据溯源标准模型,但是随着区块链的深入发展,其去中心化、可伸缩性、智能合约灵活性、溯源验证以及可信管理上的特点将被充分体现,区块链结合数据溯源管理模型将成为未来研究的重点与核心。

## 3 数据溯源计算方法

模型是溯源策略和流程的抽象表示框架,而计算则是数据溯源具体实现的技术和算法。数据溯源计算分为基于标注的溯源和基于非标注的溯源两种基本思想。前者需要对数据源和数据产生过程进行标注(如元数据、注释和声明

等),并将标注信息与数据信息一起存储和传播,溯源时通过查询目标数据的标注来获得数据的溯源信息;而后者则根据数据的转换过程进行逆向推导,不需要对源数据和目标数据进行额外标注,只在需要数据溯源时才进行计算,溯源时通过构造逆置函数进行反向查询来获得数据的溯源信息。这两种溯源思想适用于不同的应用场景,并发展出了一些具体的计算方法,本文主要介绍关系数据库、科学工作流、大数据平台、云计算和区块链等典型应用环境下溯源计算的基本方法与研究进展。

### 3.1 关系数据库下的溯源计算方法

在关系型数据库下,数据溯源计算分为三个层次:模式级溯源计算、字段级溯源计算和元组级溯源计算。模式级溯源计算目标是获取目标关系表的来源表集合及其转换过程,是一种粗粒度的数据溯源;字段级溯源计算目标是获取特定关系字段的源表字段和目标表字段在数据转换过程中的映射关系,它是模式级溯源信息的细化;元组级溯源计算目标是获取对产生目标关系元组有贡献的来源元组集合,是一种细粒度的数据溯源[57]。为了适应复杂的溯源查询,降低细粒度溯源信息的存储开销并提高溯源效率,许多研究提出了新的或优化的溯源计算方法。

Glavic 提出了 Perm 溯源方法以适应复杂的 SQL 查询,并对溯源过程进行关系表达树结构的转换描述,使得数据库查询引擎同时计算溯源数据和查询结果,较大程度上优化了计算效率,但在溯源信息转换描述时的冗余复杂度较高[58]。周忠开发了一种基于 SQL 关系代数的数据转换图方法,对涉及数据溯源的 SQL 语法进行源表、目标表以及转换节点上的高效解析及基于属性映射图的约简,但也存在只能解析标准 SQL 语法的问题[57]。在支持全特性的查询语言方面,Müller 等由程序切片(Program Slicing)和抽象解释(Abstract Interpretation)的思想派生出一种方法,将查询语言编译为一种图灵完备的抽象语言,计算和追踪细粒度的 where-和 why-溯源信息,理论上支持任何 SQL 特性,如关联子查询、递归查询、分组/聚合与关系型数据库内置函数库等[59],

Dietrich 等对支持全特性查询语言的溯源方法进一步完善,发展出了一种用于溯源的观察式 SQL 调试器,可以实时评估所选子表达式,准确地标识那些在产生可疑查询输出时的重要输入行,在最小化的输入上运行调试器产生 bug 相关的观察结果[60]。林悦邦提出了一种溯源转换的规范表示方法 PET(Provenance Explains Transformation)用以支持全特性 SQL 和过程语言,并在此基础上通过构造模式溯源有向图、字段溯源有向图及其溯源子图路径等实现各粒度溯源信息的转换过程计算[61]。

为优化细粒度溯源信息标注的存储及提高溯源计算的效率,Karvounarakis 等基于图论的思想,为数据溯源开发了专用查询语言 ProQL[62],它采用一般性的数据来源编码方式,实现来源数据在多种数据库中的低开销存储,并利用索引定制策略来保证溯源查询处理性能[63]。王黎维等基于查询结构匹配的溯源树构建溯源信息表达和存储的基本框架,避免了数据派生过程中的冗余存储,该框架下还包括一套针对关系型查询的溯源信息存储优化方法,通过查询树的部分节点来存储溯源信息的同时,在溯源信息的跟踪和计算方面均不会产生过大的时间开销[64]。Detuch 等针对 Datalog 中的溯源存储优化开展了一系列研究,在 K-Relations 基础上引入闭包构造,利用减少闭包避免数据量爆炸[65];提出一种 Top-K 查询策略,通过树模式和排序的方法选择重要的溯源信息[66];在自然语言查询接口上使用自然语言的形式扩展数据溯源,起到压缩冗余溯源信息的效果[67]。

以上数据溯源计算方法均是基于标注的思想,计算时通过与输出数据相关联的标注来确定出处。然而,在有些应用场景下,例如缺失溯源标注信息时的数据集成与数据仓库应用,需基于非标注的思想,即不存储整个过程的标注信息,而是利用少量的元数据信息实现对数据

的溯源计算。这类方法适用于高效的逆过程中,如 SPIDER 系统应用模式映射的方法,将数据溯源路径表示为目标数据所经过的变换过程,每一步变换都在模式映射的逻辑表示函数下将输入数据模式映射为输出数据模式,从而支持逆向提取溯源信息[68];Srivastava 等存储关系数据库查询相关的溯源信息的逆查询,用于实现高效的数据溯源查询[69]。但这种方法仅当存在这种已验证的逆查询机制或逆过程推导函数时才有效[70]。因此采用何种溯源方法,也需要依据数据库中数据存储策略的实际情况而定,对于非单一的溯源场景,还需将两者结合起来实现时间和空间上的均衡,这也是关系数据库溯源计算的未来发展重点。

### 3.2 科学工作流下的溯源计算方法

科学工作流广泛应用于科学研究,在科学数据生成与分析后对其进行存储、注解、归档和共享等。科学工作流溯源计算即在不同级别捕获科学数据分析处理过程,系统地获取衍生数据产品的溯源信息。在简单的科学工作流中,用户通常通过在文件名中嵌入时间戳和一些简单描述,或者通过手工方法,间接维护溯源信息。而在大体量的科学数据环境下,甚至在经过了数据库的特定管理后,面对复杂的大型数据密集型科学研究过程时,溯源过程中仍然存在如存储、效率及复杂度等性能问题[71]。为此,一些研究对科学数据工作流下的溯源展开了专门的方法研究与策略优化。

Chapman 等提出了一种分解算法以及两种基于继承的方法,用于存储科学应用所产生的复杂数据溯源信息,该方法在利用双通道的参数分解算法提高溯源计算效率的同时,使用基于结构和谓词的继承方法减少溯源数据所需的存储量[72]。王黎维等采用对象代理模型描述一系列科学任务的执行过程,使得工作流管理操作类似于传统的数据库管理操作,然后基于对象数据库中的双向指针机制,定义了一种新的数据溯源方法——双向指针追踪法[73],这一方法首先构建集成对象代理数据库的科学工作流服务框架,再利用对象代理数据库内部对象间的双向指针实现高效的数据跟踪,在特定形式的工作流数据库中供比标注溯源或逆推溯源方法更高的性能。Bao 等开发了一种可达标签的溯源查询方法,用于计算工作流中符合给定规范的来源查询,在常量时间内解决溯源可达性查询问题[74]。

针对科学工作流溯源的具体问题,也产生了一些溯源策略优化研究。吴渊在国家天文台 NebulaS 科学工作流管理系统中,增加了对嵌套工作流执行时捕获数据溯源动态信息的支持,并以 XML 格式集成于系统中[75]。邓仲华等针对原始溯源信息在支持数据发布任务时的缺陷,提出了一种提纯溯源的方法来获得适合于数据发布任务的溯源信息[76]。Billings 将分布式账本应用于科学工作流管理系统共享中的数据溯源,证明了科学工作流的溯源计算和分布式账本之间的兼容性[77]。Sarikhani 针对科学工作流系统的各类溯源信息收集机制设计了相应的优化策略[78]。魏银珍等在提出的云计算环境下科学工作流的数据溯源框架下,从原始数据、派生整合数据与文献三个不同层级收集和查询溯源信息,来提高数据溯源的保真度与灵活性[71]。

### 3.3 大数据平台下的溯源计算方法

分布式、大规模的数据集须并行计算处理,数据溯源常用于解决大数据处理中的异常数据调试、诊断和排错等问题,是对大数据平台处理功能的扩展。在溯源过程中,综合两种溯源计算思想,支持反向查询(查找对给定输出元素有贡献的输入子集)和正向查询(确定从特定输入元素派生的输出元素)。下面对面向几种主流大数据平台(Hadoop 和 Spark 等)的溯源计算方法进行介绍。

在 Hadoop 的 MapReduce 工作流下进行数据溯源是一项困难的任务,因为它采用的是批处理并行方式,一旦完成计算任务,则只留下数据集本身。为了解决这个问题,Ikeda 等[79] 和

Park 等[80]基于包装器(Wrapper)对 Hadoop 扩展了数据溯源的功能——RAMP (Reduce And Map Provenance),RAMP 可以捕获和跟踪任何 MapReduce 工作流中的溯源信息,在合理的时间和空间开销下,保留 Hadoop 的并行执行效率和容错性,同时支持高效的反向查询而无需对溯源信息进行特殊索引。Logothetis 等针对大数据分析中多步骤数据流的调整和调试,开发了一种可扩展的工具 Newt[81],用于捕获和使用元组级数据溯源来发现和解决大数据分析中的错误。Newt 的 API 支持从一系列数据密集型可扩展计算(DISC)体系结构中捕获细粒度的溯源信息以及数据挖掘策略的溯源信息等,通过多步骤、用户定义的转换进行数据流的主动记录,运行时间开销也适度。

RAMP 和 Newt 方法虽然较好地支持了 DISC 系统中数据流特别是复杂数据流的数据溯源,但也存在一些不足,如使用外部存储或分布式文件系统来保留溯源信息、在单独的编程接口中进行溯源信息查询、不支持查看中间数据或在中间数据上重现数据处理步骤等,导致了它们在特大规模数据集上的溯源效率较低,也无法支持数据的交互式调试。鉴于此,Interlandi 等在 Spark 中建成了一个交互式数据溯源的 Titian 库[82-83],直接集成 Spark 基于弹性分布式数据集 RDD (Resilient Distributed Dataset)的编程接口扩展细粒度数据溯源功能。由于 Titian 的溯源功能集成了 Spark 的内部批处理操作符和容错机制,使之能实现 Spark 的交互式终端无缝操作,最大程度减少了大规模数据集上的程序运行开销。

此外,还有一些有代表性的研究,如 Rajmohan 等针对 Spark 平台基于递归查询的溯源方案存在的空间开销过大问题,提出一种溯源图划分方法,利用工作流依赖关系图对工作流进行递归划分,减少空间开销的同时也支持大规模工作流溯源信息的有效查询[84];郝鹏飞针对 Hadoop 平台下 Oozie 模型工作流数据溯源问题[85],提出了基于有向无环图 DAG (Directed Acyclic Graph)的粗粒度数据溯源方法和基于自动标记的细粒度数据溯源方法等,从各方面不断完善大数据平台下的数据溯源研究。

### 3.4 云环境下的数据溯源计算方法

云环境中的数据溯源尤其重要,因为云中数据是全局存储的,可以广泛匿名共享。云上数据溯源的常规策略是将从多个来源收集的日志和审计数据链接在一起,以提供事件或结果的完整历史记录。但云系统由动态链接的资源组成,任何一个应用系统上的数据溯源都需要构建事件逻辑序列以审计多个来源的数据,包括应用系统本身、应用系统所使用虚拟资源(Virtual Resources)以及虚拟资源可能使用物理资源(Physical Resources)的所有日志。然后,审计员必须确定每个环节的所有可能时间间隔以及所有相关日志文件,以正确组合这些数据产品并进行溯源。考虑到云系统的潜在规模,资源收集和数据组合既不容易也不实际。因此,Abbadi 等认为云系统的所有层、子层和组都应该包含一种机制来支持可链接数据的收集,从而提供与特定活动相关的事件来源[86]。Saad 通过建立一个安全的通信通道提供对数据溯源信息的访问,并实现云中虚拟层和物理层的日志合并存储[87]。Ko 团队相继提出了基于日志分析的云数据溯源解决方案 Flogger[88]、S2Logger[89]和 Progger[90],Flogger 在云中记录以文件为中心的访问及虚拟机 VMs (Virtual Machines)和物理机 PMs (Physical Machines)间的内核空间(Kernel Spaces)传输信息,为云服务商、终端用户和监管机构提供溯源信息;S2Logger 以数据事件为中心,跨所有云服务器将文件和数据块级别捕获的数据事件链接在一起,对其进行分析并可视化生成数据事件序列,描述整个数据生命周期内的溯源信息;Progger 则进一步解决了几项关键性问题:提供日志篡改证据及防伪造功能、服务器间的多粒度时间戳精确同步及日志空间存储效率等,为数据事件的审计提供了有效保证。

除基于日志捕获和分析的云数据溯源外,Ali 扩展云计算环境下的 PROV 模型,以 PROV 模型定义根元素和边作为云 PROV 模型子类型元素和边的基础,通过增加的 5 个新类元素进行数据溯源[91]。Asghar 等根据威胁模型分析云环境下溯源信息系统的实体、潜在威胁对手和可能的攻击,提出一种支持加密搜索并同时保证云存储数据溯源保密性的方案[92]。此外,Sarwar 等采用数据包中嵌入可见水印和隐藏水印两种方法,保护云计算中共享数据对象的数据包可靠性[93]。魏银珍等设计了云计算环境下科学工作流的数据溯源框架,以及收集、存储与检索溯源信息的方法[71]。Hasham 等提出一种新的科学工作流溯源方法 ReCAP,通过比较工作流结构、所执行资源和工作流输出三个级别的溯源信息,确定云端上工作流执行的可再现性[94]。

### 3.5 区块链应用下的溯源计算方法

从应用来看,区块链可视为一个随时间顺序不断增长的去中心化的分布式数据库。区块链数据溯源的基础是信息的全局共识,包括对分布式各节点的数据验证、查询与同步,以保证区块链中纳入的数据和查询的数据均是全网公认正确的,当某个副本节点的区块中数据出错或被恶意篡改时,可通过分布式溯源修复断链点。区块链的成链就是不断把当前数据转化为历史数据的过程,而溯源就是把历史数据转化为当前数据并以所需形态呈现的过程[95]。

为保证区块链中入链数据和溯源数据准确一致,系统中的共识机制负责解决由谁来构造区块,以及如何维护区块链统一。目前成功应用于公有链、联盟链和私有链中较为典型的共识机制有工作量证明 PoW (Proof of Work)机制、实用拜占庭容错 PBFT (Practical Byzantine Fault tolerance)机制、Paxos 共识及其各类改进机制等[96]。刘家稷等将公有链和私有链一起用于存储溯源信息,应用于物联网中产品的防伪溯源系统[97]。

在完成云数据溯源任务时,云节点隐式创建的分布式网络将溯源信息记录在加密的分布式容错账本中。账本对所有节点进行更新时需取决于每个节点都同意的共识机制,但是传统共识机制具有低吐率和高延迟等缺点。Zawoad 等分析了区块链应用下云环境中可信应用程序溯源的威胁,提出了安全应用程序溯源 SECAP (Secure Application Provenance)方案,能较为有效地保证溯源所需的完整性和保密性[98]。Tosh 等构建了一个基于区块链的云计算溯源框架 BlockCloud[99],改进了 PoW 共识机制的权益证明 PoS (Proof of Stake)共识机制。Demichev 等将区块链技术、智能合约和元数据驱动型数据管理相结合,提出了一种分布式数据溯源管理方法 ProvHL (Provenance Hyper-Ledger),支持在分布式环境中存储和交换数据,并实现容错、安全的元数据溯源管理[100]。Kim 等基于形式化本体的建模方法,为物联网下产品供应链溯源的区块链运作提供了自动推理和验证的形式化规范[101]。

区块链的去中心化、不可篡改、过程透明等诸多特性,成为数据溯源的一个重要技术趋势。但区块链自身的数据查询及分析功能较为简单[56]。随着区块链平台上应用与数据规模的增大,如何更高效地管理和查询溯源信息将是区块链溯源所要解决的重要问题。

## 4 数据溯源典型应用

数据溯源在诸多领域中得到应用,产生了如 VisTrails、SPADE、Orchestra、MyGrid、CMCS 和 PASOA 等代表性的数据溯源应用系统,相关的综述论文对此也有介绍[3, 11-13]。本文主要面向信息资源管理,介绍数据溯源的典型应用及相关实践探索。

(1)数字图书馆领域的数据溯源应用。从图书馆资源建设角度来看,通过记录数字对象的溯源信息,可以为数字对象的真实性判定提供可靠依据[102]。林芳芳将 PROV 溯源模型与数字图书馆的现有元数据做关联分析[103],提出

数字图书馆处理流程的 PROV 溯源应用模型，并应用到国家科技成果转化项目库[104]中，对采集、加工、发布等流程进行规范化控制，提高了科技成果数据质量[102]。而从长期保存角度来看，数字对象在被收入到长期保存系统后会经过备份、格式迁移、内容分发等变化，将这些变化记录为溯源信息，可以为数字对象的真实性提供证据[105]。许多数字资源长期保存系统也专门支持数据溯源功能，如佛罗里达图书馆自动中心开发的 DAITSS(Dark Archive In The Sunshine State)系统[106]和中国科学院文献情报中心开发的数字资源长期保存系统(DPS)[105, 107]等。此外，姚敏设计了一个基于区块链的图书馆数据仓储，对图书馆数据更新流程、区块链交易数据保存与溯源的策略等进行了分析[108]。

(2)档案信息管理中的数据溯源应用。当前，数字档案大多通过数字水印和数字签名来进行保护，其真实性主要依赖于系统的权限信任，一旦发生系统入侵或内部授权用户的非法篡改，则会直接破坏档案数据的真实性[109]。为此，2017 年和 2018 年国家档案局科技项目立项选题指南提出在电子文件归档与电子档案管理方面应用区块链技术[110]。在应用层面，王平等提出了基于区块链的电子文件可信保护框架[111]；谭海波通过构建数字档案馆间联盟链和公有链的结合方式，利用智能合约和 IPFS(星际文件系统)集群加密等技术，实现数字档案的保护、验证、恢复与共享，目前已初步应用于中科院合肥研究院档案馆(下属多个研究所形成联盟)[109]。但是，也有研究认为目前实现区块链档案管理缺少必要的环境，也没有成熟的经验可借鉴[112]。

(3)网络信息资源管理中的数据溯源应用。倪静等基于 PROV 模型创建了在线论文的溯源描述[113]；贾君枝等使用 PROV 模型运用维基条目生命周期编纂活动的页面记录生成溯源信息，并实现了溯源信息的表示、存储、可视化及查询等操作[114]；张子良等构建了微博的 PROV 数据溯源模型[115]；何冰心基于社区结构与人群行为特性模拟信息在社交网络中的传播模式，在 Pinto 等人提出的基于观察点部署的信息溯源模型[116]的基础上，加入时间和距离相关性的溯源规则，提高信息溯源定位的准确率[117]；彭朝君[118]、宋磊[119]、李城等[120]使用数据溯源方法研究了复杂社交网络中的谣言传播、判别及溯源问题；Maemura 等讨论了网络信息归档的收集策略与溯源记录框架[121]。此外，开放数据的可追溯性一直以来备受关注，我国十几个地方政府的数据开放网站，其元数据中包含溯源元数据[122-123]。

(4)科学数据共享管理中的数据溯源应用。科学数据的共享和再利用是科技创新和知识发现的重要驱动因素之一[124]，许多国家建设了科学数据存储与共享平台，如英国国际数据存档 UKDA、澳大利亚研究数据存储基础设施 RDSI 与美国康奈尔大学的 DataStaR 等[125]。我国也陆续建成了国家基础科学数据共享服务平台[126]、国家人口与健康科学数据共享服务平台[127]及南京大学人文社科大数据平台[128]等，但平台在数据准确性、数据用途追踪与数据提供者权利保障等方面还不够理想[129]。郝世博等构建了科学数据共享层次化区块链架构，探索其交互信息、数据区块、共识机制、智能合约等实现机理[124]。谷俊等依托 Hyperledger Fabric 构建了人文社科数据共享联盟链模型[129]。

(5)电子商务信息系统中的数据溯源应用。食品供应链上的每个环节都可能发生数据断裂与信息篡改，从而产生虚假的产品信息[130]。为此，政府常强制上下游企业使用政府主导开发的商品溯源系统，如加拿大国家农业和食品安全追溯系统(National Agriculture and Food Traceability System, NAFTS)[131]、我国的食品工业企业质量安全追溯平台[132]等。京东基于区块链开发了数据交易溯源平台“京东万象”，赋予商品“数据可溯源”标识，将交易信息加密存储在区块链中[133]；天猫国际利用区块链技术、药监码技术以及大数据跟踪进口商品全链路，给进口商品打上唯一的身份证码，将商品生产、检测、运输、通关等环节的信息完整展现在用户面

前[134]。此外,区块链企业 VeChain 以构造可信任分布式商业生态环境为宗旨,开发了一站式区块链 BaaS(Blockchain as a Service)服务平台 ToolChain[135],结合物联网设备,在产品全生命周期的关键节点采集信息到区块链存证,提供商品溯源防伪、供应链全流程管控、数据认证与流程认证等一系列区块链综合服务[136]。

## 5 数据溯源发展展望

数据溯源的目的是为用户提供当前数据的历史档案,通过溯源展现数据从产生到最终状态的整个演变过程,确保数据的安全、可靠和完整,增强用户对数据的信任。目前,关于数据溯源的研究虽有一定进展,但也面临不少挑战,未来还需在如下方面深入研究:

(1)在数据高速增长与动态变化的情形下,优化与创新发展与之相适的溯源信息捕获、存储、查询和表示技术。随着数据特别是移动端数据的爆发式增长和高频率更新,捕获到的溯源信息可能比原始数据本身的体量还要大。为此,不仅要考虑大存储空间和低存储成本,还要考虑数据处理能力、请求响应能力、环境评估和效益分析等问题。另外,溯源信息关联分析的难度影响了溯源信息的检索效率,例如通过本体映射方式构建数据的关联关系,数据一旦发生变化,原有关联关系就不再适用,需重新进行分析。如何在溯源信息的关联关系发生变化时最小化计算代价,是一个迫切需要解决的问题。此外,简洁易懂的溯源信息呈现形式至关重要。目前数据库溯源信息的呈现形式通常为有向无环图,虽然比较直观,但复杂数据的溯源图过于"笨重",还需寻找更优雅、美观、高效的溯源信息表示方法。

(2)在分布式异构数据环境下,加强溯源信息的访问接口规范、结构化转换技术及可信计算方法的标准化研究。现阶段大多数数据溯源模型和框架都是针对特定任务而构建,即便是通用的溯源模型也在不同应用领域进行了继承和扩展,所开发出的数据溯源管理系统大多是在一个独立的系统内部实现溯源信息的管理与利用。跨系统溯源信息的访问和获取以基于包装器的 API 为主,扩展性和通用性不足。如何建立起一套标准的访问接口规范是亟待解决的问题之一。而且,多样化的数据溯源模型和溯源信息格式,导致溯源信息的大规模异构现象,影响检索效率和共享效果。例如,同一系统内的数据溯源信息可能来自不同传感器设备,格式存在差别,难以使用统一的检索标准;不同系统间溯源信息更是存在 XML、日志文本和数据表等多种形式,阻碍了溯源信息的共享和互操作。针对此类问题,除了进一步研发结构化转换技术外,还需要注重顶层设计,逐步建立溯源数据处理的技术标准。

(3)发展基于多方法融合的数据溯源信息安全技术,提升溯源信息安全保障能力。溯源信息本身的机密性、隐私性、不可篡改性、不可伪造性和不可否认性是数据溯源应用推广的基本条件。首先,数据溯源信息经常比数据对象本身更机密,比如一项军事科研工作的溯源信息会展示其流转过程中每一环节、每名参与人员、每个仪器设备等所有细节;其次,溯源信息经常会涉及个人隐私,当数据与来自多个端点的元数据一起被收集、整理和分析时,隐私性往往难以保证;第三,具有凭证意义的溯源信息要求不可篡改,但不法分子受到利益驱动时可能会对溯源信息进行非法操作,破坏其原始性和完整性;第四,不可伪造性要求溯源信息内容真实可靠,既要防止伪造,又要确保即使伪造了数据也能检测出来;最后,不可否认性是指溯源信息一旦被记录下来,用户就不能否认之前采取的任何操作处理,问题发生时可以据此进行问责。此外,云计算环境下动态数据的匿名、广泛共享特点,也对数据溯源信息的安全性提出了更高的要求。在这样的要求下,依靠单一的数字签名、密码校验、访问控制及区块链技术无法满足实际需求,势必需要进行多方法融合的数据溯源安全技术研究。

(4)充分利用区块链的技术优势,多方面探

索和完善基于区块链的数据溯源模型、方法及应用。区块链通过分布式账本技术和共识机制实现去中心化,确保溯源数据的完整性和不可篡改性;通过智能合约降低数据的收集成本并提高数据的应用效率,保证了数据的高质量和有效性。当前,一些基于区块链的数据溯源研究与实践也暴露出了一些具体问题。例如无法支持复杂场景下可信数据的有效管理,大规模链式结构或日志结构上的溯源信息查询还相当低效;特别地,区块链作为一种创新技术,在很多具体领域尚无特别有效的应用场景和成功经验,甚至在有些领域还与现有系统的管理技术、手段和制度发生冲突。但是,许多区块链溯源研究充分证明了区块链的应用价值。随着技术的成熟,数据溯源的应用场景将会不断扩展,不但可以验证与扩展已有的标准、模型和方法,也可以更多地解决现实问题,如政府数据治理、共享与开放,等等。

## 参考文献


[1] Foster I, Wilde M, Zhao Y, 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. Washington, D.C, USA, 2002: 37-46.

[2] 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.)

[3] 明华, 张勇, 符小辉. 数据溯源技术综述[J]. 小型微型计算机系统, 2012, 33(9): 1917-1923. (Ming Hua, Zhang Yong, Fu Xiaohui. Survey of data provenance[J]. Journal of Chinese Mini-Micro Computer Systems, 2012, 33 (9): 1917-1923.)

[4] Medeiros N. A public trust: libraries and datacuration[J]. OCLC Systems & Services: International Digital Library Perspectives, 2013, 29(4): 192-194.

[5] 王芳, 慎金花. 国外数据管护研究与实践进展[J]. 中国图书馆学报, 2014, 40(4): 116-128. (Wang Fang, Shen Jinhua. Advances in data curation abroad: research and practice[J]. Journal of Library Science in China, 2014, 40 (4): 116-128.)

[6] 陈骥. 建立健全食品安全溯源体系的思考[J]. 中国市场监管研究, 2011(6): 49-52.(Chen Ji. Thoughts on establishing and perfecting food safety provenance system[J]. Research on China Market Regulation, 2011 (6): 49-52.)

[7] Muniswamy-Reddy K K. Foundations for provenance-awaresystems[D]. Cambridge: Harvard University, 2010.

[8] 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, 64. (Li Xiumei, Wang Fengying. Research on data provenance's security model[J]. Journal of Shandong University of Technology(Natural Science Edition), 2010, 24 (4): 56-60,64.)

[9] Simmhan Y L, Plale B, Gannon D. A survey of data provenance techniques[EB/OL]. [2019-06-20]. http://www.cs.indiana.edu/pub/techreports/TR618.pdf.

[10] 刘喜平, 万常选. 数据起源研究综述[J]. 科技广场, 2005(1): 47-52. (Liu Xiping, Wan Changxuan. Research on data provenance: an overview[J]. Science Mosaic, 2005 (1): 47-52.)

[11] 戴超凡, 王涛, 张鹏程. 数据起源技术发展研究综述[J]. 计算机应用研究, 2010, 27(9): 3215-3221. (Dai Chaofan, Wang Tao, Zhang Pengcheng. Survey of data provenance technique[J]. Application Research of Computers, 2010, 27(9): 3215-3221.)

[12] 高明, 金澈清, 王晓玲, 等. 数据世系管理技术研究综述[J]. 计算机学报, 2010, 33(3): 373-389. (Gao

Ming, Jin Chengqing, Wang Xiaoling, et al. A survey on management of data provenance[J]. Chinese Journal of Computers, 2010, 33 (3): 373-389.)

[13] 徐扬, 王申罡. 数据起源研究进展[J]. 情报理论与实践, 2016, 39(7): 136-140. (Xu Yang, Wang Shenquan. Progress of data provenance research[J]. Information Studies: Theory & Application, 2016, 39 (7): 136-140.)

[14] Herschel M, Diestelkämper R, Lahmar H B. A survey on provenance: what for? what form? what from?[J]. The VLDB Journal, 2017, 26(6): 881-926.

[15] 石丽波, 孙连山, 王艺星. 数据起源安全研究综述[J]. 计算机应用研究, 2017, 34(1): 1-7. (Shi Libo, Sun Lianshan, Wang Yixing. Survey of data provenance security[J]. Application Research of Computers, 2017, 34 (1): 1-7.)

[16] Wang Y R, Madnick S E. A polygen model for heterogeneous database systems: the source tagging perspective [C]// Proceedings of the 16th International Conference on Very Large Data Bases. San Francisco, California, USA, 1990: 519-538.

[17] Lanter D P. Design of a lineage-based meta-data base for GIS[J]. Cartography and Geographic Information Systems, 1991, 18(4): 255-261.

[18] Cui Y, Widom J, Wiener J L. Tracing the lineage of view data in a warehousing environment[J]. ACM Transactions on Database Systems (TODS), 2000, 25(2): 179-227.

[19] Buneman P, Khanna S, Tan W C. Data provenance: some basic issues[C]//Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science. New Delhi, India, 2000, 1974: 87-93.

[20] Buneman P, KhannaS , Tan W C. Why and where: a characterization of data provenance[C]// Proceedings of the 17th International Conference on Data Engineering. London, UK, 2001: 316-330.

[21] Goble C. Position statement: musings on provenance, workflow and (Semantic Web) annotations for bioinformatics[C]// Proceedings of Workshop on Data Derivation and Provenance. Chicago, Illinois, USA, 2002: 1-5.

[22] Greenwood M, Goble C, Stevens R, et al. Provenance of e-science experiments-experience from bioinformatics [C]// Proceedings of the 2nd UK e-Science All Hand Meeting. Nottingham, UK, 2003: 223-226.

[23] Glavic B, Dittrich K R. Data provenance: a categorization of existing approaches[C]// Datenbank systeme in Business, Technologie Und Web (BTW 2007). Aachen, Germany, 2007: 227-241.

[24] Braun U, Shinnar A, Seltzer M. Securing provenance[C]// Proceedings of 3rd USENIX Workshop on Hot Topics in Security. San Jose, CA, USA, 2008.

[25] Cheney J, Chiticariu L, Tan W C. Provenance in databases: why, how, and where[J]. Foundations and Trends in Databases, 2009, 1(4): 379-474.

[26] Ram S, Liu J. Understanding the semantics of data provenance to support active conceptual modeling[C]// Active Conceptual Modeling of Learning, Next Generation Learning-Base System Development. Tucson, Arizona, USA, 2006:17-29.

[27] 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.)

[28] 倪静. 语义 Web 环境下基于模型的数据溯源研究[D]. 北京: 中国农业科学院, 2014. (Ni Jing. Model-based data provenance in semantic web environment [D]. Beijing: Chinese Academy of Agricultural

Sciences, 2014.)

[29] Moreau L, Clifford B, Freire J, et al. The open provenance model core specification (v1.1)[J]. Future Generation Computer Systems, 2011, 27(6): 743-756.

[30] Zhao J. Open provenance model vocabulary specification[EB/OL]. [2019-06-20].http://open-biomed.sourceforge.net/opmv/ns.html.

[31] Groth P, Moreau L. PROV-overview: an overview of the PROV family of documents[EB/OL]. [2019-06-20]. https://www.w3.org/TR/prov-overview/.

[32] Moreau L, Missier P. PROV-DM: the PROV datamodel[EB/OL]. [2019-06-20]. https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

[33] 陈希, 胡良霖, 朱艳华,等. 数据溯源描述模型国家标准研制与推广[J]. 标准科学, 2019(4): 108-112. (Chen Xi, Hu Lianglin, Zhu Yanhua, et al. Research and popularization of national standard for data provenance descriptive model[J]. Standard Science, 2019 (4): 108-112.)

[34] IPAW 2008[EB/OL]. [2019-06-20]. http://www.vldb.org/dblp/db/conf/ipaw/ipaw2008.html.

[35] Wang M, Blount M, Davis J, et al. A time-and-value centric provenance model and architecture for medical event streams[C]// Proceedings of the 1st ACM Sigmobile International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments. San Juan, Puerto Rico, 2007:95-100.

[36] 陈颖. 一种基于DNA双螺旋结构的数据起源模型[J]. 现代图书情报技术, 2008(10): 11-15. (Chen Ying. A data provenance model based on the double-helical structure of DNA[J]. New Technology of Library and Information Service, 2008 (10): 11-15.)

[37] Kulkarni D. A provenance model for key-value systems[C]// Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance. Lombard, Illinois, 2013:12.

[38] 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.)

[39] 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 spproach[D]. Wuhan: Wuhan University, 2013.)

[40] Hartig O, Zhao J. Publishing and consuming provenance metadata on the web of linked data[C]// Proceedings of the 3rd International Provenance and Annotation Workshop(IPAW 2010). Troy, New York, USA, 2010: 78-90.

[41] Cyganiak R, Zhao J, Alexander K, et al. Vocabulary of interlinked datasets (VoID)[EB/OL]. [2019-06-20]. http://vocab.deri.ie/void#Dataset.

[42] 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, 29(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, 29 (2): 18-23.)

[43] Belhajjame K, Cheney J, Corsar D, et al. PROV-O: the PROV ontology[EB/OL]. [2019-06-20]. https://www.w3.org/TR/prov-o/.

[44] Hasan R, Sion R, Winslett M. Introducing secure provenance: problems and challenges[C]// Proceedings of the 2007 ACM Workshop on Storage Security and SurVivability (StorageSS). Alexandria, Virginia, USA, 2007: 13-18.

[45] Hasan R, Sion R, Winslett M. The case of the fakepicasso: preventing history forgery with secure provenance [C]// Proccedings of the 7th USENIX Conference on File & Storage Technologies. San Francisco, California, USA, 2009: 1-14.

[46] Zhang J, Chapman A, Lefevre K. Do you know where your data's been? -tamper-evident database provenance [C]// Proceedings of the 6th VLDB Workshop on Secure Data Management. Lyon, France, 2009:17-32.

[47] 李秀美. 数据起源安全模型研究[D]. 淄博: 山东理工大学, 2011. (Li Xiumei. Research on data provenance's security model[D]. Zibo: Shandong University of Technology, 2011.)

[48] Park J, Nguyen D, Sandhu R. A provenance-based access control model[C]// Proceedings of the 2012 Tenth Annual International Conference on Privacy, Security and Trust (PST). Washington, D.C, USA, 2012: 137-144.

[49] 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(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 (6): 6-11.)

[50] 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 provenance security model [J]. Research on Library Science, 2016 (20): 36-41.)

[51] 袁勇, 王飞跃. 区块链技术发展现状与展望[J]. 自动化学报, 2016, 42(4): 481-494. (Yuan Yong, Wang Feiyue. Blockchain: the state of the art and future trends [J]. Acta Automatica Sinica, 2016, 42 (4): 481-494.)

[52] Ramachandran A, Kantarcioglu M. Smart Provenance: a distributed, blockchain based data provenance system [C]// CODASPY '18 Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy. Tempe, Arizona, USA, 2018: 35-42.

[53] Neisse R, Steri G, Fovino N I. A blockchain-based approach for data accountability and provenance tracking[J]. arXiv preprint arXiv: 1706.04507, 2017.

[54] 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(11A): 367-368, 381. (Liu Yaozong, Liu Yunheng. Security provenance model for RFID big data based on blockchain[J]. Computer Science, 2018, 45 (11A): 367-368,381.)

[55] Liang X, Shetty S, Tosh D, et al. ProvChain: a blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability[C]// The 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (IEEE/ACM CCGrid 2017). Madrid, Spain, 2017: 468-477.

[56] 钱卫宁, 邵奇峰, 朱艳超,等. 区块链与可信数据管理:问题与方法[J]. 软件学报, 2018, 29(1): 150-159. (Qian Weining, Shao Qifeng, Zhu Yanchao, et al. Research problems and methods in blockchain and trusted data management[J]. Journal of Software, 2018, 29 (1): 150-159.)

[57] 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.)

[58] Glavic B. Perm: efficient provenance support for relational databases[D]. Zurich: University of Zurich, 2010.

[59] Müller T, Grust T. Provenance for SQL through abstract interpretation: value-less, but worthwhile[C]// Proceedings of the 41st International Conference on Very Large Databases (VLDB 2015). Kohala Coast, Hawaii, USA, 2015: 1872-1875.

[60] Dietrich B, Müller T, Grust T. The best bang for your bu(ck)g-when SQL debugging and data provenance go hand in hand[C]// Proceedings of the 19th International Conference on Extending Database Technology (EDBT). Bordeaux, France, 2016: 674-675.

[61] 林悦邦. 面向关系数据库的数据起源研究与设计[D]. 广州：华南理工大学，2017.（Lin Yuebang. A research and implementation of provenance for relational databases[D]. Guangzhou：South China University of Technology，2017.）

[62] Karvounarakis G，Ives Z G，Tannen V. Querying data provenance[C]// Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indianapolis，Indiana，USA，2010：951-962.

[63] Green T J，Karvounarakis G，Tannen V. Provenance semirings[C]// Proceedings of the 26th ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems. Beijing，China，2007：31-40.

[64] 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.）

[65] Deutch D，Milo T，Roy S，et al. Circuits for datalog provenance[C]// Proceedings of International Conference on Database Theory. Athens，Greece. 2014：201-212.

[66] Deutch D，Gilad A，Moskovitch Y. Selective provenance for datalog programs using top-k queries[C]// Proceedings of the 41st International Conference on Very Large Data Bases. Kohala Coast，Hawaii，USA，2015，8(12)：1394-1405.

[67] Deutch D，Frost N，Gilad A. NLProv：natural language provenance[C]// Proceedings of the VLDB Endowment. New Delhi，India，2016，9(13)：1537-1540.

[68] Alexe B，Chiticariu L，Tan W C. SPIDER：a schema mapping debugger[C]// Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul，Korea，2006：1179-1182.

[69] Srivastava D，Velegrakis Y. Intensional associations between data and metadata[C]// Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing，China，2007：401~412.

[70] 夏菁. 基于可信度计算的不确定数据起源研究[D]. 南京：南京航空航天大学，2011：17-18.（Xia Jing. Research of uncertain data provenance based on confidence computing[D]. Nanjing：Nanjing University of Aeronautics and Astronautics，2011：17-18.）

[71] 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）：115-118，114.（Wei Yinzhen，Deng Zhonghua. Research on provenance data collection and query framework of scientific workflow in cloud[J]. Information Studies：Theory & Application，2015，38（7）：115-118，114.）

[72] Chapman A，Jagadish H V，Ramanan P V. Efficient provenance storage[C]// Proceedings of the 2008 ACM SIGMOD international conference on Management of Data. Vancouver，Canada，2008：993-1006.

[73] 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.）

[74] Bao Z，Davidson S B，Milo T. Labeling workflow views with fine-grained dependencies[J]. Proceedings of the VLDB Endowment. Istanbul，Turkey，2012，5(11)：1208-1219.

[75] 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.）

[76] 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.)

[77] Billings J J. Applying distributed ledgers to manage workflow provenance[J]. arXiv preprint arXiv: 1804.05395, 2018.

[78] Sarikhani M, Wendelborn A. Mechanisms for provenance collection in scientific workflow systems[J]. Computing, 2018, 100(5): 439-472.

[79] Ikeda R, Park H, Widom J. Provenance for generalized map and reduce workflows[C]// The 5th Biennial Conference on Innovative Data Systems Research (CIDR 2011). Asilomar, California, USA, 2011: 273 - 283.

[80] Park H, Ikeda R, Widom J. RAMP: a system for capturing and tracing provenance in mapreduce workflows[C]// The 37th International Conference on Very Large Data Bases (VLDB 2011). Seattle, Washington, USA, 2011: 1351-1354.

[81] Logothetis D, De S, Yocum K. Scalable lineage capture for debugging DISC analytics[C]// Proceedings of the 4th annual Symposium on Cloud Computing. Santa Clara, California, USA. 2013: 17.

[82] Interlandi M, Shah K, Tetali S D, et al. Titian: data provenance support inspark[J]. Proceedings of the VLDB Endowment. Kohala Coast, Hawaii, 2015, 9(3): 216-227.

[83] Interlandi M, Ekmekji A, Shah K, et al. Adding data provenance support to apache spark[J]. The VLDB Journal, 2018, 27(5): 595-615.

[84] Rajmohan C, Lohia P, Gupta H, et al. Efficiently processing workflow provenance queries on spark[J]. arXiv preprint arXiv: 1808.08424v2, 2018.

[85] 郝鹏飞. 大数据模型分析平台下的数据溯源关键技术研究[D]. 成都: 电子科技大学,2017:4-16. (Hao Pengfei. Research on the key technology of data provenance based on big data model analysis platform[D]. Chengdu: University of Electronic Science and Technology, 2017:4-16.)

[86] Abbadi I M, Lyle J. Challenges for provenance in cloud computing[C]// Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance. Heraklion, Crete, Greece, 2011:1-6.

[87] Saad M IM , Jalil K A , Manaf M . Data provenance trusted model in cloud computing[C]//Proceeding of the 2013 International Conference on Research and Innovation in Information Systems (ICRIIS). Kuala Lumpur, Malaysia, 2013: 257-262.

[88] Ko R K L, Jagadpramana P, Lee B S. Flogger: a file-centric logger for monitoring file access and transfers within cloud computing environments[C]// Proceedings of the 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications. Washington, D.C, USA, 2011: 765-771.

[89] Suen C H, Ko R K L, Tan Y S, et al. S2Logger: end-to-end data tracking mechanism for cloud data provenance [C]// Proceedings of the 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Washington, D.C, 2013: 594-602.

[90] Ko R K L, Will M A. Progger: an efficient, tamper-evident kernel-space logger for cloud data provenance tracking [C]// Proceedings of the 2014 IEEE International Conference on Cloud Computing. Washington, D.C, USA, 2014: 881-889.

[91] Ali M. Provenance-based data traceability model and policy enforcement framework for cloud services[D]. Southampton: University of Southampton, 2016.

[92] Asghar M R, Ion M, Russello G, et al. Securing data provenance in the cloud[C]// Proceedings of the 2011 IFIP WG 11.4 international conference on Open Problems in Network Security. Berlin, Heidelberg, 2012: 145-160.

[93] Sarwar M U, Hanif M K, Talib R, et al. Data provenance for cloud computing using watermark[J]. International

Journal of Advanced Computer Science & Applications, 2017, 8(6): 407-411.

[94] Hasham K, Munir K. Reproducibility of scientific workflows execution using cloud-aware provenance (ReCAP)[J]. Computing, 2018, 100(12): 1299-1333.

[95] 张力. 基于区块链的数据溯源研究[D]. 贵州: 贵州大学, 2018. (Zhang Li. Research on data provenance based on blockchain[D]. Guizhou: Guizhou University, 2018.)

[96] Zheng Z, Xie S, Dai H, et al. An overview of blockchain technology: architecture, consensus, and future trends[C]//Proceedings of the 2017 IEEE International Congress on Big Data (BigData Congress). Honolulu, Hawaii, USA, 2017: 557-564.

[97] 刘家稷, 杨挺, 汪文勇. 使用双区块链的防伪溯源系统[J]. 信息安全学报, 2018, 3(3): 17-29. (Liu Jiaji, Yang Ting, Wang Wenyong. Traceability system using public and private blockchain[J]. Journal of Cyber Security, 2018, 3 (3): 17-29.)

[98] Zawoad S, Hasan R . SECAP: towards securing application provenance in the cloud[C]//Proceedings of the 2016 IEEE International Conference on Cloud Computing. San Francisco, California, USA, 2016: 900-903.

[99] Tosh D K, Shetty S, Liang X, et al. Consensus protocols for blockchain-based data provenance: challenges and opportunities[C]//Proceeding of the 2017 IEEE 8th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference. New York, USA, 2017: 469-474.

[100] Demichev A, Kryukov A, Prikhodko N. The approach to managing provenance metadata and data access rights in distributed storage using the hyperledger blockchain platform[C]//2018 Ivannikov Ispras Open Conference (ISPRAS).2018.

[101] Kim H M, Laskowski M. Towards an ontology-driven blockchain design for supply chain provenance[J]. Social Science Electronic Publishing, 2016.

[102] 林芳芳. PROV 溯源模型在数字图书馆的应用研究[D]. 北京: 中国科技信息研究所, 2016. (Lin Fangfang. The application research of PROV model in Digital Library[D]. Beijing: Institute of Scientific and Technical Information of China, 2016.)

[103] 林芳芳, 赵辉. 数据溯源模型与 DC 元数据的映射研究[J]. 数字图书馆论坛, 2016(3): 10-16. (Lin Fangfang, Zhao Hui. Study on mapping between data provenance model an DC metadata[J]. Digital Library Forum, 2016 (3): 10-16.)

[104] 国家科技成果转化项目库[EB/OL]. [2019-06-09]. https://www.nstad.cn/. (National Science and Technology Achievements Database [EB/OL]. [2019-06-09]. https://www.nstad.cn/.)

[105] 吴振新, 李文燕, 蒋世银. 基于 OAIS 信息模型的起源信息封装策略研究与应用[J]. 图书情报工作, 2017(18): 113-118. (Wu Zhenxin, Li Wenyan, Jiang Shiyin. Research and application of the provenance information packaging strategy based on the OAIS information model[J]. Library and Information Service, 2017 (18): 113-118.)

[106] DAITSS[EB/OL]. [2019-06-09]. http://daitss.fcla.edu/.

[107] 吴振新, 李文燕, 蒋世银. 构建以事件为核心的长期保存系统起源管理框架[J]. 图书情报工作, 2016(6): 91-96. (Wu Zhenxin, Li Wenyan, Jiang Shiyin. Construction of a provenance management framework for long-term preservation system based on the event-centric[J]. Library and Information Service, 2016 (6): 91-96.)

[108] 姚敏. 基于区块链技术的图书馆科学数据仓储模型研究[J]. 图书馆学刊, 2018(6):94-97. (Yao Min. Research on library science data warehousing model based on blockchain technology[J]. Journal of Library Science, 2018 (6): 94-97.)

[109] 谭海波, 周桐, 赵赫,等. 基于区块链的档案数据保护与共享方法[J/OL]. 软件学报,2019: 1-15[2019-

06-11]. https://doi.org/10.13328/j.cnki.jos.005770. (Tan Haibo, Zhou Tong, Zhao He, et al. Archives data protection and sharing method based on blockchain [J/OL]. Journal of Software,2019: 1-15 [2019-06-11]. https://doi.org/10.13328/j.cnki.jos.005770.)

[110] 国家档案局. 通知公告[EB/OL]. [2019-06-11]. http://www.saac.gov.cn/daj/tzgg/list.shtml. ( National archives Administration of China. Notice[EB/OL]. [2019-06-11]. http://www.saac.gov.cn/daj/tzgg/list.shtml.)

[111] 王平, 李沐妍, 姬荣伟. 基于区块链技术的电子文件可信保护框架研究[J]. 档案学研究, 2019(1): 101-107. (Wang Ping, Li Muyan, Ji Rongwei. Research on the trusted protection framework for electronic records based on blockchain technology[J]. Archives Science Study, 2019 (1): 101-107.)

[112] 李高峰, 马国胜, 胡国强. 现阶段区块链技术在档案管理中不可行分析[J]. 档案管理, 2018(5): 30-32. (Li Gaofeng, Ma Guosheng, Hu Guoqiang. Feasibility analysis of blockchain technology in archives management [J]. Archives Management, 2018 (5): 30-32.)

[113] 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.)

[114] 贾君枝, 刘超. 维基数据条目编纂起源模型构建[J]. 情报理论与实践,2019, 42(6): 71-26. (Jia Junzhi, Liu Chao. Construction of provenance model about wikidata item compilation [J]. Information Studies: Theory & Application,2019, 42(6): 71-76.)

[115] 张子良, 董红斌, 谭成予,等. 微博信息可信度评估的数据起源方法[J]. 计算机应用研究, 2018, 35(11): 3330-3334. (Zhang Ziliang, Dong Hongbin, Tan Chengyu, et al. Evaluation of weibo credibility based on data provenance[J]. Application Research of Computers, 2018, 35 (11): 3330-3334.)

[116] Pinto P C, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks[J]. Physical Review Letters, 2012, 109(6): 068702-5.

[117] 何冰心. 社交网络中信息传播和溯源模型研究[D]. 大连: 大连理工大学, 2018. (He Bingxin. A study on information diffusion and source locating model in social network [D]. Dalian: Dalian University of Technology, 2018.)

[118] 彭朝君. 基于复杂社交网络的谣言传播和溯源模型[D]. 成都: 西南交通大学, 2016. (Peng Chaojun. Rumor spread and traceable models of complex social network [D]. Chengdu: Southwest Jiaotong University, 2016.)

[119] 宋磊. 社交网络中流言的判别与溯源[J]. 电脑知识与技术, 2019, 15(4): 6-8. (Song Lei. Discrimination and traceability of rumors in social networks[J]. Computer Knowledge and Technology, 2019, 15 (4): 6-8.)

[120] 李城, 沙俊淞, 武文. 基于最长公共子序列的微博谣言溯源研究[J]. 计算机与现代化, 2018(1): 107-112. (Li Cheng, Sha Junsong, Wu Wen. Research on origin of micro-blog rumors based on longest common subsequence[J]. Computer and Modernization, 2018 (1): 107-112.)

[121] Maemura E, Worby N, Milligan I, et al. If these crawls could talk: studying and documenting web archives provenance[J]. Journal of the Association for Information Science and Technology, 2018, 69(10):1223-1233.

[122] 陈红玉, 翟军, 袁长峰,等. 开放政府数据的溯源元数据研究及应用[J]. 情报杂志, 2017(6): 148-155. (Chen Hongyu, Zhai Jun, Yuan Changfeng, et al. Provenance metadata research of open government data and application discussion[J]. Journal of Intelligence, 2017 (6): 148-155.)

[123] 黄如花, 王春迎. 我国政府数据开放平台现状调查与分析[J]. 情报理论与实践, 2016, 39(7): 50-55. (Huang Ruhua, Wang Chunying. Investigation and analysis of the status of open government data platform in China[J]. Information Studies: Theory & Application, 2016, 39 (7): 50-55.)

[124] 郝世博, 邓雨亭. 融合数据监管与数据溯源的科学数据共享管理研究[J]. 情报理论与实践, 2018, 41

(3): 23-28. (Hao Shibo, Deng Yuting. Research on scientific data sharing governance integrating data curation and data provenance[J]. Information Studies: Theory & Application, 2018, 41 (3): 23-28.)

[125] 完颜邓邓, 高峰. 英美澳科学数据存储与共享平台建设现状调查及启示[J]. 图书馆建设, 2016(3): 29-34. (Wanyan Dengdeng, Gao Feng. Study on the status and enlightenment of scientific data storage and sharing platform in England, America and Australia[J]. Library Development, 2016 (3): 29-34.)

[126] 国家基础科学数据共享服务平台[EB/OL]. [2019-06-20]. http://www.nsdata.cn/. (National basic science data sharing service platform[EB/OL]. [2019-06-20]. http://www.nsdata.cn/.)

[127] 国家人口与健康科学数据共享服务平台[EB/OL]. [2019-06-20]. http://www.bmicc.cn/. (National science data sharing platform for population and health[EB/OL]. [2019-06-20]. http://www.bmicc.cn/.)

[128] 南京大学人文社科大数据平台[EB/OL]. [2019-06-20]. http://skbigdata.nju.edu.cn/data/. (Big data platform for humanities and social sciences of Nanjing University[EB/OL]. [2019-06-20]. http://skbigdata.nju.edu.cn/data/.)

[129] 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.)

[130] 肖程琳, 李姝萱, 胡敏思,等. 区块链技术在食品信息溯源中的应用研究[J]. 物流工程与管理, 2018, 40(8):77-79. (Xiao Chenglin, Li Shuxuan, Hu Minsi, et al. Application of blockchain technology in traceability of food information[J]. Logistics Engineering and Management, 2018, 40 (8):77-79.)

[131] Government of Canada. Agriculture and agri-food Canada[EB/OL]. [2019-06-20]. http://www.agr.gc.ca/eng/home/?id=1395690825741.

[132] 工业和信息化部消费品工业司.食品工业企业质量安全追溯平台[EB/OL]. [2019-06-20]. http://www.foodcredit.org.cn/zs/zhuisu.html. (Department of Consumer Goods Industry, Ministry of Industry and Information Technology of People's Republic of China. Food industry enterprise quality and safety traceability platform[EB/OL]. [2019-06-20]. http://www.foodcredit.org.cn/zs/zhuisu.html.)

[133] 数据溯源平台使用文档[EB/OL]. [2019-06-20]. https://wx.jdcloud.com/help/center/48. (Data provenance platform manual[EB/OL]. [2019-06-20]. https://wx.jdcloud.com/help/center/48.)

[134] 区块链在天猫国际商品溯源中的应用[EB/OL]. [2019-06-20]. https://yq.aliyun.com/articles/348787?utm_content=m_40006. (Application of blockchain in provenance of tmall international commodity [EB/OL]. [2019-06-20]. https://yq.aliyun.com/articles/348787?utm_content=m_40006.)

[135] Vechain[EB/OL]. [2019-06-20]. https://www.vechain.com/.

[136] 全球4大区块链溯源项目追踪[EB/OL]. [2019-06-20]. https://www.wanlianzhijia.com/News/show/id/33948.html. (Tracking of the four global blockchain provenance projects [EB/OL]. [2019-06-20]. https://www.wanlianzhijia.com/news/show/id/33948.html.)

**王 芳** 南开大学商学院信息资源管理系教授,博士生导师。天津 300071。
**赵 洪** 南开大学商学院信息资源管理系博士研究生。天津 300071。
**马嘉悦** 南开大学商学院信息资源管理系硕士研究生。天津 300071。
**李晓阳** 南开大学商学院信息资源管理系硕士研究生。天津 300071。
**张晓玥** 南开大学商学院信息资源管理系硕士研究生。天津 300071。

(收稿日期:2019-06-25;修回日期:2019-08-20)