·专题一:区块链技术及应用·

区块链与数据治理

孟小峰* 刘立新

(中国人民大学 信息学院,北京 100872)

[摘 要] 当下,大数据的“堰塞湖”已经形成,数据治理问题迫在眉睫。传统的治理概念来自政府、企业、IT 领域,数据治理既有其一般性,也有其特殊性。本文提出数据治理的根本保障在于增加大数据价值实现过程的透明性。区块链凭借去中心、公开透明和不可篡改的特性与大数据价值实现的透明性需求相契合,能够克服当前数据治理存在的问题,为数据治理提供了新的解决思路。同时,基于区块链实现数据治理也面临诸多挑战。

[关键词] 数据治理;区块链;隐私保护;溯源问责;决策可信

大数据时代,数据源源不断产生并自主汇聚至多方数据收集者,数据已经成为企业间竞争的关键和影响国家竞争力的重要因素,由此数据治理成为企业治理和国家治理的重点领域和重要方式[1,2]。然而,大规模数据收集也带来严峻的隐私泄露、数据滥用和数据决策不可信等问题,对传统的数据治理提出了新的挑战。例如,“Facebook-剑桥分析”事件[3]就是大规模数据收集导致的隐私泄露、数据滥用和决策不可信的典型案例。进一步,大规模数据自主汇聚还导致数据垄断困境的出现,使数据被不合理的分配与享用[4]。大数据的“堰塞湖”已经产生,如何使这些问题得到有效解决,并使数据得到正确和规范的使用是决定大数据继续发挥价值的关键,也是目前数据治理亟待解决的问题。

上述问题产生的主要原因是大数据价值实现过程的不透明。大数据收集和共享流通过程不透明导致隐私泄露和数据滥用等问题追踪问责困难,并且致使数据垄断问题悄然形成却缺乏评估和解决依据;大数据存储、处理和共享流通等过程中缺乏透明导致数据被篡改等问题难以被发现,影响决策数据质量并最终导致数据决策不可信。由此可以得出,当前数据治理的根本保障在于增加大数据价值实现过程的透明性。数据收集和共享流通过程透明地对数据流向进行记录,以溯源问责的方式进行隐私保护[5]和为解决数据垄断提供依据;数据存储、处理和共享流通等过程透明使决策数据可审计和促进数据决策可信。数据治理实现途径有多种方式,除了法律法规和政策标准,还需要技术方法的保驾护航。区块链起源于数字货币,具有公开透明、去中心和不可篡改的特性。该技术的进步发展为解决当前数据治理面临的问题带来新的机遇[6-10]。

孟小峰 博士,中国人民大学教授,博士生导师,CCF 会士,主要研究方向为数据库理论与系统、大数据管理系统、大数据隐私保护、大数据融合与智能、大数据实时分析、社会计算等。

本文提出了数据治理的根本保障在于增加大数据价值实现过程中的透明性,总结了数据治理的发展历程和技术上实现数据治理的关键内容,并对基于区块链实现数据治理的研究现状进行分析和总结,最后提出目前数据治理面临的挑战。

1 数据治理概述

“治理”(Governance)一词起源于拉丁文“掌舵”(Steering),最初用于“政府治理”,目标是协调政府与其他社会主体之间的利益。后来逐渐受到企业的

收稿日期:2019-12-30;修回日期:2020-02-13

* 通信作者,Email:xfmeng@ruc.edu.cn

本文受到国家自然科学基金项目(91646203、61532010、91846204、61532016)的资助。

认同和重视，出现了“企业治理”，目标是协调企业内部利益相关者的利益。伴随着 IT 资源和数据资源的日益丰富，又出现了“IT 治理”和“数据治理”[1, 2]。后来，由于大数据的流通性、多源数据融合和涉及多方参与主体等应用特性，“数据治理”又进一步延伸，出现了“大数据治理”。“大数据治理”关注大数据生命周期中数据生产者、数据收集者、数据使用者、数据处理者和数据监管者①等各方参与主体，其目标是在兼顾各方参与主体的权利、责任和利益的前提下发挥数据价值，即大数据价值实现和风险规避。

由于“大数据治理”是“数据治理”的延伸，为避免混淆，本文后续内容采用“数据治理”的概念来探讨大数据时代的数据治理。数据治理的发展过程和涉及的参与主体如图 1 所示。

大数据的应用特性与数据治理的目标决定了当下数据治理的关键内容。目前，数据治理的关键内容和挑战聚焦在以下 3 个方面：

（1）提高决策数据质量。大数据价值实现需要多源数据的融合，然而大数据来源广泛且生命周期内涉及多方参与主体，数据是否真实产生、数据被篡改和多源数据的标准和类型不一致等问题都会影响决策数据质量，进而影响数据使用者的数据决策结果。所以，数据治理需要支持大数据在其全生命周期内的溯源。

（2）评估与监管个人隐私数据的使用。大数据应用的流通特征使数据生产者对数据获取和共享缺乏知情权和控制权。作为数据生产者，用户不知道哪些数据被收集、被谁收集、收集之后流向哪里和作何使用。同时，数据的收集汇聚导致数据垄断现象出现。数据垄断可能会阻碍市场竞争、使消费者福利受损、阻碍行业技术创新和带来更严重的个人隐私泄露风险等问题，但数据监管者却无法对数据应用进行评估和监管；此外，大数据应用的多源数据融合特征还可能会引发更严峻的隐私泄露问题。所以，数据治理需要对个人隐私数据使用进行评估与监管。

（3）促进数据共享。数据共享可以促进大数据价值实现和缓解数据垄断，但同时也需要解决隐私保护等问题。一方面，数据共享双方之间发生数据共享流通时，考虑到隐私问题，需要以有效的方式保护数据生产者的个人隐私。另一方面，限于法律和实际应用中的一些因素，需要在不直接传输原始数据情况下，依据多方数据持有者的数据实现分布式数据集进行统计分析和分布式机器学习。由于多方参与者之间不存在完全的可信性，此时应该能够保护数据使用者对其共享过程进行验证。所以，数据治理需要在权衡数据生产者和数据使用者等参与主体利益的前提下促进数据共享。

数据治理需要综合法律法规、政策标准和技术方法等多种途径实现。一方面，国际组织和国家相关部门出台相应的法律法规和政策标准。例如，国际数据治理研究所从组织、规则和过程三方面总结数据治理的要素[11]；以及，国际标准 ISO/IEC 38505-1：《信息技术—IT 治理—数据治理》为数据治理参与主体提供原则、定义以及模型，帮助数据治理参与主体评估、指导和监督其数据利用的过程[12]。另一方面，数据治理亟需安全、可靠的技术方法，为大数据应用过程中数据隐私保护、提高决策

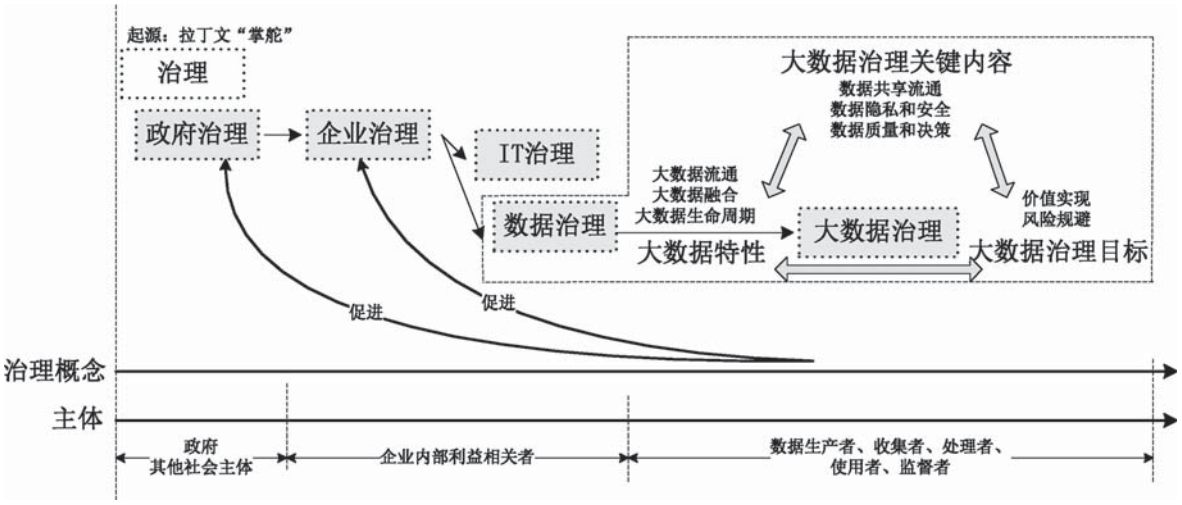

图 1　数据治理发展过程和涉及的参与主体

① 各参与主体之间可能存在重合，例如当数据收集者自己使用数据并且具有处理能力时，数据收集者也充当数据处理者和数据使用者。

数据质量、促进数据共享和评估监管数据应用的合规性等问题提供技术支持。

2 基于区块链实现数据治理

区块链本质上是一种去中心化的分布式数据库，在增加大数据价值实现过程的透明性方面具有天然的优势，为解决当前数据治理的关键问题提供了可行性。

2.1 支持审计的数据存储和处理

数据决策渗透在人们生产、生活的方方面面，由于涉及多方利益相关者，数据在存储、处理和共享流通等过程中存在数据被篡改、数据伪造，以及不同来源数据的类型和标准规则差异等问题，这些问题都会影响决策数据质量。所以，数据使用者需要对决策数据进行审计。区块链作为去中心化的分布式数据库，可以实现支持审计的数据存储和处理。此外，基于区块链在不同利益主体之间构建去中心分布式数据库系统，数据通过全网快速广播至各个利益主体，也能够保证数据共享流通的真实性和及时性。

区块链网络内各节点都存储数据，数据一旦存入区块链就不会被篡改或者丢失，即使存在通信故障和蓄意攻击等问题，也仍然能保证数据存储的正确性，数据使用者可以对其进行审计。此外，将数据存入区块链还支持数据处理过程和处理结果的可审计性。对于传统的数据库管理系统，数据库中存储和维护当前数据状态，仅将数据处理过程等信息存在数据库日志，用于故障恢复，并不支持数据的历史状态查询。然而，区块链作为去中心分布式数据库，支持数据的历史状态查询，用以确认当前数据状态是否正确。基于区块链进行数据存储和处理，在保险[13]、医疗[14-17]和供应链[18-21]等数据完整性要求较高领域是有重要意义的。由此，数据使用者可以对决策数据进行审计并在可信数据上执行分析和进行决策[22-25]。

针对不同来源数据的类型和标准规则不一致等问题，可以基于区块链和智能合约制定统一的数据类型和标准规则。智能合约会被存储和同步在区块链各个节点，区块链会根据智能合约上的代码自动执行验证。由于智能合约的执行过程公开透明，使其执行过程和执行结果是可审计的，能提高多源数据共享效率且不存在单点失败。

2.2 支持溯源问责的数据获取和共享

在传统的数据获取和数据共享过程，由数据收集者制定数据使用协议并据此告知用户数据收集、共享和使用等信息。用户作为数据生产者，对数据的知情权和可控权仍然限于法律约束和第三方信用背书。然而，由于数据获取和共享等过程对外不可见，其契约履行情况也无从考证。2014 年皮尤研究中心关于美国隐私状况的报告指出，91％的受访者认为他们已经失去对数据收集者收集和使用个人数据的控制，61％的受访者对不了解数据收集者如何使用个人数据感到沮丧[26]；2016 年《中国网民权益保护调查报告》显示，84％的网民对个人隐私泄露带来的不良影响有深切的感受[27]。数据获取和数据共享不透明导致隐私泄露问题更为严峻。传统的加密、差分等隐私保护技术虽然对数据隐私具有一定的保护作用，但是目前还不足以应对大规模数据收集带来的隐私泄露风险。应用区块链的去中心性和不可篡改性，可以记录数据的获取和共享情况，进一步实施追踪溯源，并结合策略承诺（Policy Compliance）、违反检测（Violation Detection）和隐私审计（Privacy Audit），可以在隐私保护技术无效的情况下以溯源问责的方式保护隐私，也可以为评估监管数据和解决数据垄断问题提供技术支持。

目前，已有研究利用区块链增加移动应用[28]、医疗[29, 30]和物联网[31-33]等领域的数据获取和共享流通的透明性。基于区块链实现数据获取和共享的框架可以分为四层：数据获取层—存储层—区块链层—共享层。在数据获取层，数据生产者对数据收集内容、形式和目的等具有知情权；在存储层，采用传统数据库管理系统、云存储和分布式存储系统等方式存储数据，并采用加密技术对数据进行加密来保护数据安全和隐私；在区块链层，由区块链执行去中心化的访问控制，使任何数据访问情况都通过区块链的交易被记录在区块链；在共享层，实现数据共享并对共享关系进行保护。正是通过上述四层，区块链增加数据获取和共享流通的透明性。

2.3 支持验证的分布式数据统计分析和机器学习

在医学研究、公共安全和商业合作等一些应用领域，需要在大规模分布式数据集上执行统计分析[34-36]和机器学习任务[38-41]，但考虑法律法规等因素的限制，需要在不泄露隐私数据前提下进行分布式数据统计分析和机器学习。针对分布式数据集统计分析，现有方案基于安全多方计算、秘密共享、本地化差分隐私和同态加密等技术实现。然而，安全多方计算方法不适用于大规模数据提供者参与；秘密共享使数据提供者失去数据控制权；本地化差分隐私需要平衡数据的可用性和隐私损失；同态加密

能够保证数据提供者不失去数据控制权，而且不需要考虑隐私损失，但是实现的前提是数据提供者提供真实数据和计算节点的可信计算。针对分布式机器学习，由于数据提供者和数据需求者之间不存在完全的信任，各个数据提供者也可能会提供不可靠的数据或参数扰乱最终结果，以及由于经济利益等因素提前退出。所以，数据使用者需要对分布式数据集统计分析和分布式机器学习进行验证，以及需要合理的经济激励促进其顺利执行。

基于区块链实现可验证的分布式数据集统计分析常包括数据提供者、多个计算节点、多个验证节点和数据查询者。其中，数据提供者提供加密数据，多个结算节点执行密文计算，由区块链组成多个验证节点并对计算节点的计算进行验证。除此之外，分布式数据集统计分析需要考虑数据机密性、数据提供者和数据之间不可连接性、查询结果机密性和计算结果的鲁棒性等安全和隐私问题。为此通常采用洗牌和同态加密等技术进行保护。

基于区块链实现可验证的和公平的分布式机器学习，数据提供者将本地机器学习参数上传和存储至区块链，由区块链执行交叉验证，将分布式机器学习过程的每一步都记录在区块链。同时，还可以结合零知识证明和密码学承诺对恶意的参与方进行经济惩罚，通过经济激励促进公平。除此以外，分布式机器学习需要考虑数据提供者本地参数的安全性，因为本地参数也可能会泄露数据或者机器学习模型。为此通常采用差分隐私、秘密共享和同态加密等技术对其进行保护。

3　挑战与问题

区块链为数据治理提供了新的思路，但数据治理具体实现过程中也将面临诸多挑战，同时对区块链自身技术有了更高的要求。此外，基于区块链实现数据治理会导致政府和企业的管控机制和业务流程发生重大变革，这将对政府管理和企业管理提出新挑战。目前，数据治理实现过程面临的挑战与问题主要包括以下 3 个方面：

（1）数据治理实现过程中面临的挑战。一方面，虽然将数据共享流通信息记录在区块链可以实现溯源问责，但是在大规模数据收集和数据共享流通错综复杂背景下，如何实现跨平台和跨领域的溯源问责是具有挑战性的问题。同时，溯源问责也可能会带来隐私泄露问题，所以溯源问责过程的隐私保护也至关重要。另一方面，虽然将数据存入区块链，可以一定程度上防止数据篡改和保证数据可以进行追踪溯源，但是保证数据存入区块链之前的真实性和可靠性仍存在挑战。

（2）对区块链自身技术提出的新挑战。区块链自身的存储需求限制、隐私与安全、可扩展性和互操作性等方面还存在大量待解决的问题，现有比特币、以太坊和超级账本等主流的区块链还不能满足数据治理的需求。为此应该考虑设计轻量级的、高可扩展的、互联通性较强的适用于数据治理需求的区块链。同时，伴随着各类区块链系统的出现，区块链系统评价标准与评估规范也成为亟待解决的问题。

（3）对政府管理和企业管理提出的挑战。区块链的去中心化特性将打破传统的中心化管理方式，对政府和企业的管理权威带来挑战；同时，去中心化特性还会使数据安全和保密的责任置于多方，对政府和企业的数据管理等方面带来新的挑战。此外，基于区块链实现数据治理并据此对数据执行相应的监管措施需要一个过程，而且随着区块链技术的迅猛发展，将会对传统的监管制度和法律法规政策提出新的要求。

4　结　语

数据治理已经成为国家治理和企业治理的重点领域和重要因素。随着各个领域数据的不断开放共享，数据治理对数据共享、数据监管和隐私保护等方面都提出了更高的要求。这些问题通过与区块链相结合可以提升数据治理的效率和透明度，将会有利于构建一个全新的数据信息时代。与此同时也会带来诸多新的挑战，需要多学科、多领域和多部门共同的努力去实现数据治理的新篇章。

参　考　文　献

[1] 吴信东，董丙冰，堵新政，等. 数据治理技术. 软件学报，2019，30(9)：2830—2856.

[2] 安小米，郭明军，魏玮，等. 大数据治理体系：核心概念、动议及其实现路径分析. 情报资料工作，2018，(1)：5—11.

[3] Jennifer Zhu Scott. Facebook and Cambridge Analytica: what you need to know as fallout widens. https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html. [2018-03-19]/[2020-01-01].

[4] 孟小峰，朱敏杰. 数据垄断与其治理模式研究. 信息安全研究，2019，1(9)：789—797.

[5] 孟小峰，张啸剑. 大数据隐私管理. 计算机研究与发展，2015，52(2)：265—281.

[6] 祝烈煌,高峰,沈蒙, 等. 区块链隐私保护研究综述. 计算机研究与发展, 2017, 54(10): 2170—2185.

[7] 袁勇, 倪晓春, 曾帅, 等. 区块链共识算法的发展现状与展望. 自动化学报, 2018, 44(11): 93—104.

[8] 邵奇峰, 金澈清, 张召, 等. 区块链技术:架构及进展. 计算机学报, 2018, 41(5): 3—22.

[9] 韩璇, 袁勇, 王飞跃. 区块链安全问题:研究现状与展望. 自动化学报, 2019, 45(1): 208—227.

[10] 李芳, 李卓然, 赵赫. 区块链跨链技术进展研究. 软件学报, 2019, (6): 1649—1660.

[11] The Data Governance Institute. data governance institute framework. http://www.datagovernance.com/wp-content/uploads/2014/11/dgi_framework.pdf. [2014-11-15]/[2020-02-13].

[12] 国家标准化管理委员会.《信息技术—IT 治理—数据治理—第 1 部分:ISO/IEC 38500 在数据治理中的应用》. http://www.sac.gov.cn/sgybzeb/gzdt_2132/201705/t20170515_238441.htm. [2017-05-15]/[2020-02-13].

[13] Vo H. Blockchain-based data management and analytics for micro-insurance applications//Proc of the ACM Int Conf on Information and Knowledge Management. New York: ACM, 2017: 2539—2542.

[14] Vo, H. Research directions in blockchain data management and Analytics//Proc of Int Conf on Extending Database Technology. Bordeaux: Springer LNCS, 2018: 445—448.

[15] Vo H. Blockchain-Powered big data analytics platform// Proc of the Int Conf on Big Data Analytics. Berlin: Springer, 2018: 15—32.

[16] Shae Z, Tsai J P. On the design of a blockchain platform for clinical trial and precision medicine// Proc of the Int Conf on Distributed Computing Systems. Washington: IEEE, 2017: 1972—1980.

[17] Tsai J. Transform blockchain into distributed parallel computing architecture for pecision medicine//Proc of the Int Conf on Distributed Computing Systems. Washington: IEEE, 2018: 1290—1299.

[18] Xu XW, Lu QH, Liu Y. Designing blockchain-based applications a case study for imported product traceability. Future Generation Computer Systems 2019, 92: 399—406.

[19] Swan M. Blockchain: Blueprint for a new economy // O'Reilly Media Inc, 2015: 1—18.

[20] Vasco L, Luís A. An overview of blockchain integration with robotics and artificial intelligence [EB/OL]. arXiv preprint, arXiv: 1810.00329, 2018[2018-09-30]. https://arxiv.org/abs/1810.00329

[21] Salah K, Rehman MHU, Nizamuddin N, et al. Blockchain for AI: review and open research challenges. IEEE Access, 2019, 7: 10127—10149.

[22] Li Y, Zheng K, Yan Y. EtherQL: A query layer for blockchain system// Proc of the Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2017: 556—567.

[23] Xu C, Zhang C, Xu J. vChain: Enabling verifiable boolean range queries over blockchain databases [EB/OL]. arXiv preprint, arXiv: 1812.02386, 2018[2018-12-06]. https://arxiv.org/abs/1812.02386.

[24] Zhang C, Xu C, Xu J, et al. GEM^2-Tree: A gas-efficient structure for authenticated range queries in blockchain// Proc of the 35th Int Conf on Data Engineering. Washington: IEEE, 2019: 842—853.

[25] P Ruan, Chen G, TTA Dinh. Fine-grained, secure and efficient data provenance on blockchain systems//Proceeding of the Very Large DataBase. California: ACM, 2019: 975—988Explainable artificial intelligence: A survey.

[26] Pew Research Center. Public perceptions of privacy and security in the post-Snowden era. https://www.pewinternet.org/2014/11/12/public-privacy-perceptions/. [2019-01-30]/[2020-01-01].

[27] 中国互联网协会.《中国网民权益保护调查报告 2016》. http://www.isc.org.cn/zxzx/xhdt/listinfo-33759.html. [2016-06-26]/[2020-01-01].

[28] Zyskind G, Nathan O. Decentralizing privacy: using blockchain to protect personal data// Proc of IEEE Security and Privacy Workshops. Washington: IEEE, 2015: 180—184.

[29] Azaria A, Ekblaw A, Vieira T. MedRec: using blockchain for medical data access and permission management// Proc of the Int Conf on Open & Big Data. Washington: IEEE, 2016: 25—30.

[30] Dubovitskaya A, Xu Z, Ryu S. Secure and trustable electronic medical records sharing using blockchain. American Medical Informatics Association., 2017, 650—659.

[31] Ouaddah A, Abou Elkalam A, Ait Ouahman A. FairAccess: a new blockchain-based access control framework for the Internet of Things. Security and Communication Networks, 2016, 9(18): 5943—5964.

[32] Hossein S, Lukas B. Droplet: Decentralized authorization for IoT Data Streams [EB/OL]. arXiv preprint, arXiv: 1806.02057, 2018 [2018-11-14]. https://arxiv.org/abs/1806.02057.

[33] Li R, Song T, Mei B. Blockchain for large-scale internet of things data storage and protection. IEEE Transactions on Services Computing, 2018: 1—8.

[34] Henry C, Dan B. Prio: Private, robust, and scalable computation of aggregate statistics// Proc of the 14th USENIX Symposium on Networked Systems Design and Implementation, Berkeley CA: USENIX, 2017: 259—282.

[35] Froelicher D, Egger P. UnLynx: a decentralized system for privacy-conscious data sharing// Proc on Privacy Enhancing Technologies. NJ: IEEE, 2017: 232—250.

[36] Froelicher D, Juan R. Drynx: Decentralized, secure, verifiable system for statistical queries and machine learning on distributed datasets [EB/OL]. arXiv preprint, arXiv: 1902. 03785, 2019 [2019-02-11]. https://arxiv. org/abs/1902. 03785.

[37] Nelson Kibichi Bore, Ravi Kiran Raman. Promoting distributed trust in machine learning and computational simulation via a blockchain network. http://arxiv. org/abs/1810. 11126.

[38] Ravi K, Roman V, Michael H. Trusted multi-party computation and verifiable simulations: a scalable blockchain approach [EB/OL]. arXiv preprint, arXiv: 1809. 08438, 2018[2018-09-22]. https://arxiv. org/abs/1809. 08438.

[39] Tsung T, Lucila O. ModelChain: decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks [EB/OL]. arXiv preprint, arXiv:1802. 01746, 2018[2018-02-06]. https://arxiv. org/abs/1802. 01746.

[40] Weng J, Zhang J. Deepchain: auditable and privacy-preserving deep learning with blockchain-based incentive. Cryptology ePrint Archive, Report 2018/679.

[41] KUO, Tsung-Ting; GABRIEL, Rodney A, et al. Fair compute loads enabled by blockchain: sharing models by alternating client and server roles. Journal of the American Medical Informatics Association, 2019, 26(5): 392—403.

Blockchain and Data Governance

Meng Xiaofeng　　Liu Lixin

(*School of Information, Renmin University of China, Beijing* 100872)

Abstract　The "Quake lake" of big data has been formed, and the data governance is one of the most pressing problems that we face. The concept of "governance" comes from the "government governance", "enterprise governance" and "IT governance". This paper proposed that the fundamental guarantee of data governance lies in increasing the transparency of the data lifecycle. Blockchain provides a new solution for data governance by virtue of its characteristics of decentralization, transparency and tamper-proof, which can overcome the existing problems of data governance. At the same time, implementing data governance based on blockchain also faces many challenges.

Keywords　data governance; blockchain; privacy protection; traceability; accountability

(责任编辑　齐昆鹏)