引用格式： 朱艳华, 胡良霖, 廖方宇, 等.《科学数据溯源元数据》国家标准研制与实践[J].标准科学, 2025(10):93-98.
ZHU Yanhua,HU Lianglin,LIAO Fangyu,et al. Development and Practices of the National Standard for Scientific Data Provenance Metadata [J].Standard Science,2025(10):93-98.

《科学数据溯源元数据》国家标准研制与实践

朱艳华[1,2] 胡良霖[1,2*] 廖方宇[1] 高瑜蔚[1,2] 赫运涛[3] 王志强[4]
（1.中国科学院计算机网络信息中心; 2.国家基础学科公共科学数据中心; 3.国家科技基础条件平台中心; 4. 中国标准化研究院）

摘 要：【目的】为应对科学数据动态演变带来的质量评价和语义描述挑战，亟须制定科学数据溯源元数据标准，发挥数据溯源模型的积极作用。【方法】全面梳理GB/T 43707—2025《科学数据溯源元数据》研制过程，分析标准设计框架，总结标准核心内容，跟踪标准实践与推广情况。【结果】该标准可显著增强科学数据的描述能力和互操作能力，为科学数据质量评价提供了可操作的溯源规范。【结论】GB/T 43707—2025的颁布实施填补了我国数据管理标准领域空白，通过规范科学数据溯源模型描述信息，为数据质量验证、安全治理及共享重用提供可靠依据，对推动科学数据管理与应用具有重要意义。

关键词： 科学数据；数据溯源；国家标准；标准实践
DOI编码： 10.3969/j.issn.1674-5698.2025.10.013

Development and Practices of the National Standard for Scientific Data Provenance Metadata

ZHU Yanhua[1,2] HU Lianglin[1,2*] LIAO Fangyu[1] GAO Yuwei[1,2]
HE Yuntao[3] WANG Zhiqiang[4]
(1. Computer Network Information Center, Chinese Academy of Sciences; 2. National Basic Science Data Center;
3. National Science and Technology Infrastructure Center; 4. China National Institute of Standardization)

Abstract: [Objective] To address the challenges in quality assessment and semantic description arising from the dynamic evolution of scientific data, it is imperative to establish a scientific data provenance metadata standard to effectively leverage the advantages of data provenance models. [Methods] This study systematically reviews the development process of GB/T 43707-2025 *Scientific data provenance metadata*, analyzes its design framework, summarizes core components, and tracks its implementation and dissemination progress. [Results] The national standard demonstrates significant improvements in enhancing scientific data descriptive capability and interoperability while providing operational provenance specifications for scientific data quality assessment. [Conclusion] The promulgation and implementation of GB/T 43707-2025 fills a critical gap in China's data management standardization system. By standardizing scientific data provenance model descriptions, it establishes a reliable foundation for data quality verification, security governance, and sharing/reuse, thereby making substantial contributions to advancing scientific data management and applications.

Keywords: scientific data; data provenance; national standards; standard practices

---

基金项目： 本文受中国科学院院长基金特别支持项目（项目编号：E3292301）资助。
作者简介： 朱艳华，硕士，高级工程师，研究方向为大数据技术与标准规范，数据应用服务。
胡良霖，通信作者，硕士，正研级高工，硕士研究生导师，研究方向为大数据技术与标准规范、数据应用服务。

1　标准研制背景

科学数据作为一类重要的数据，已经成为国际科学研究和发展的战略性、基础性资源，是国际竞争的重要领域[1]。与此同时，人工智能领域取得的重大突破也充分彰显数据的关键作用。大量的科学和工程实践表明：只要找到足够多具有代表性的样本（数据），就可以运用数据找到一个模型或者一组模型的组合，使其和真实情况非常接近[2]。无论是基座模型，还是领域模型，在其训练与微调的过程中均离不开大量高质量、高价值密度的数据支撑[3]。然而，当前高质量数据资源匮乏的现状已成为我国在相关领域获取技术领先优势的瓶颈之一。

随着国家对数据质量重视程度不断提升，构建完善的数据质量评价体系的需求也日益凸显。2018年发布的 GB/T 36344—2018《信息技术 数据质量评价指标》[4]，尽管已经提出了规范性、完整性、准确性等基础性评价指标，但其采用的静态评价模式在面对科学数据动态演变的复杂特性时显得力不从心，难以有效适应现实状况。为了更加全面地对科学数据质量进行验证，亟须引入过程追溯机制，使数据质量评价体系能够契合科学数据的实际需求。

数据溯源（Data Provenance）作为一种重要的数据管理手段，其核心价值在于能够通过翔实记录数据的产生、转换及传播的全过程，为验证数据的真实性和有效性提供证据链。得益于语义网的快速发展，国际上先后提出了开放溯源模型OPM[5]、Provenir模型[6]、Linked Data溯源模型[7]，探索定义数据溯源模型的不同组件与实体。我国于2017年发布GB/T 34945—2017《信息技术 数据溯源描述模型》，构建了一种兼具灵活性与轻量级特点的通用描述框架，从理论层面为数据溯源工作提供规范基础[8]。不过，在实际应用过程中，该模型却面临着模型抽象度高、语义描述信息不统一等问题，导致其在跟踪数据生存周期活动中难以发挥应有作用，进而影响数据溯源方案在实际场景中的落地实施效果。

当前数据溯源模型语义描述信息主要依赖以下两种途径[9]：（1）基于数据应用程序自动对数据处理过程进行记录；（2）基于数据提供者所发布的元数据信息提取相应的记录。然而，在实际操作过程中，往往只有少量经过处理的溯源信息能够被应用程序及时记录下来。通过人工提交完整的溯源元数据信息成为落实数据溯源工作较为可行的一种方案。现有科学数据元数据标准大多侧重于对数据内容的描述，针对溯源活动及执行实体等方面缺乏规范要求。这种结构性信息的缺失现状使得科学数据溯源难以实现端到端的可信追溯，极大地制约了数据溯源工作在数据质量、数据安全等方面发挥作用。

在此背景下，2022年12月《科学数据溯源元数据》（计划号20221229–T–306）正式获批立项。2025年1月，国家标准化管理委员会发布《中华人民共和国国家标准公告（2025年第2号）》，GB/T 43707—2025《科学数据溯源元数据》正式发布并同步实施[10]。该标准从数据集、溯源活动及执行实体三大维度出发，构建科学数据溯源元数据框架，并对科学数据通用领域溯源元数据的格式和内容进行了明确的界定，为其他学科领域依据自身实际情况开展相应的扩展和补充提供参考，在改善数据质量、提升数据安全等方面发挥积极的作用。

2　标准研制过程

《科学数据溯源元数据》国家标准是我国在科学数据治理领域的一项自主创新成果。该标准由中华人民共和国科学技术部提出，并由全国科技平台标准化技术委员会（SAC/TC 486）归口管理。在标准研制过程中，组建了一支汇聚多方力量的参编团队，涵盖8个国家科学数据中心及8家科研院所、高等院校和行业龙头企业，共计16家核心参编单位。这些单位通过长达5年的协同合作与攻坚克难，共同推动了标准制定工作的顺利完成。

从主要起草单位看，包含中国科学院计算机网

络信息中心、国家科技基础条件平台中心、广州物联网研究院、中国标准化研究院、国家海洋信息中心、国家气象信息中心、中国农业科学院农业信息研究所、中国科学院地理科学与资源研究所、中国林业科学研究院资源信息研究所、中国科学院过程工程研究所、中国科学院微生物研究所、中路高科交通科技集团有限公司、中国科学院青藏高原研究所、自然资源部第一海洋研究所、南方电网互联网服务有限公司、北京信息科技大学。这些重要机构在标准研制过程中发挥了关键作用。

在标准研制过程中，起草组围绕元数据标准、数据溯源描述模型、科学数据生存周期特征、科学数据质量等关键问题，开展了系统性的研究工作。为扎实推进标准研制工作，起草组采用文献调研、工作组讨论、实地走访及专家研讨等多种方式，攻克数据溯源存在的诸多难点问题，界定了科学数据溯源元数据的核心框架和内容模块，为标准内容的科学性与合理性奠定了基础。

具体来说，《科学数据溯源元数据》主要研制过程历经了标准预研、标准立项、标准征求意见、标准审查、标准报批和标准发布实施6个关键阶段。详见表1。

3 标准主要内容

3.1 标准框架

《科学数据溯源元数据》构建了一套完整的元数据架构体系，如图1所示。该体系采用三维度建模方法描述数据溯源信息。具体包括：（1）数据集元数据维度描述数据实体的基本信息与状态特征，涵盖数据标识信息和内容特征描述等方面；（2）活动元数据维度完整记录科学数据全生存周期的关键节点，包含采集加工、存储备份、传输交换、开放共享、使用服务和安全处置等过程；（3）执行实体元数据维度详细描述每个活动的参与主体，涵盖人员角色（责任人/参与者等）和技术工具

表1《科学数据溯源元数据》国家标准主要研制过程

序号	阶段名称	主要内容
1	标准预研	2020年1月，标准起草组正式成立。起草组以中国科学院计算机网络信息中心作为牵头单位，同时有国家科技基础条件平台中心、中国标准化研究院、国家气象信息中心等多家单位共同参与
		2021年1月，标准起草组选取国家微生物科学数据中心和国家生态科学数据中心作为实地调研对象，开展调研活动
		2021年10月，标准起草组完成标准草案，并依规报送至全国科技平台标准化技术委员会（SAC/TC 486）
2	标准立项	2022年3月，《科学数据溯源元数据》参加国家标准立项答辩，评审专家同意该标准通过立项。起草组根据专家意见进一步修改标准内容
		2022年12月，标准起草组接到《科学数据溯源元数据》国家标准的制定任务，计划编号为20221229–T–306
		2023年1月，全国科技平台标准化技术委员会（SAC/TC 486）秘书处组织召开标准启动会
3	标准征求意见	2023年1～4月，标准起草组面向社会公开征求意见，同时向20个国家科学数据中心、31个国家生物种质与实验材料资源库、全国信息安全标准化技术委员会（SAC/TC 260）定向征求意见。根据反馈意见，对标准文本进行修改，形成标准送审稿
4	标准审查	2023年5月，全国科技平台标准化技术委员会（SAC/TC 486）召开标准审查会，对本文件的制定工作程序和主要技术内容进行审查。标准起草组根据意见对送审稿进行修改，形成标准报批稿
5	标准报批	2023年9月，标准起草组进一步修改标准报批稿及其他报批文件，提交审核，等待报批
6	标准发布实施	2025年1月，国家标准化管理委员会发布了《中华人民共和国国家标准公告（2025年第2号）》，《科学数据溯源元数据》正式发布并同时实施

（软件系统/硬件设备等）。

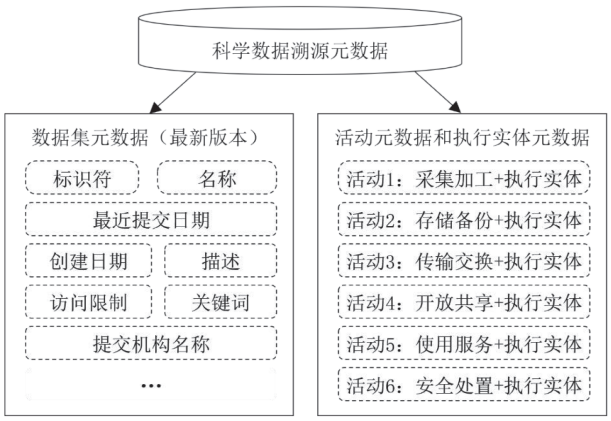

图1 科学数据溯源元数据构成示意图

《科学数据溯源元数据》在科学数据治理领域具有重要的实践价值。首先，通过元数据记录实现数据操作行为的全流程追踪，确保过程透明化；其次，为数据质量问题的根源分析提供丰富的信息支持，建立质量问题与操作环节的关联映射；最后，明确各参与主体的操作痕迹与责任边界，提升整体安全防护水平。该标准的实施将显著提升科学数据规范化管理水平，为科研诚信建设、数据共享交换以及跨领域协作提供坚实的技术基础。

3.2 标准描述方法与编写规范

《科学数据溯源元数据》依据GB/T 30523—2023《科技资源核心数据》所规定的摘要表示方式，对溯源元数据进行定义和描述。每个元数据元素涵盖了中文名称、定义、英文名称、字段类型、值域、短名和注解等关键要素，确保溯源元数据描述的系统性与规范性。

在制定过程中，《科学数据溯源元数据》参考和引用了一系列的相关标准，这些标准在不同维度为其提供重要的依据与支撑。具体包括：GB/T 7408—2005《数据元和交换格式 信息交换 日期和时间表示法》（已废止）、GB/T 32843—2016《科技资源标识》、GB/T 5271.1—2000《信息技术词汇 第1部分：基本术语》、GB/T 18391.1—2009《信息技术 元数据注册系统（MDR）第1部分：框架》、GB/T 19710—2005《地理信息 元数据》（已废止）、GB/T 30522—2014《科技平台 元数据标准化基本原则与方法》、GB/T 30523—2023《科技资源核心元数据》、GB/T 33674—2017《气象数据集核心元数据》、GB/T 34945—2017《信息技术 数据溯源描述模型》、GB/T 43708—2025《科学数据安全要求通则》及 GB/T 43705—2025《科学数据安全分类分级指南》等。通过借鉴标准的相关内容与规范要求，进一步完善标准体系和标准内容。

《科学数据溯源元数据》在整个制定过程中，严格按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定开展编写工作，切实保障标准的编写质量，使其符合标准化工作的专业要求。从标准的主要技术内容来看，其与现行国家标准和行业标准的相关规定保持高度契合，对科学数据通用溯源元数据格式和内容予以规范，并为学科领域开展相应的扩展和补充提供了可靠的参照依据，有助于推动科学数据溯源工作在不同学科领域的推广实施。

3.3 标准核心内容

《科学数据溯源元数据》国家标准规定了科学数据溯源元数据的总体要求、数据集元数据、活动元数据、执行实体元数据，并给出了元数据示例，适用于科学数据溯源过程的跟踪与记录。依据其内在逻辑与功能特性，该标准将科学数据溯源元数据的核心字段划分为3个主要部分，即数据集元数据、活动元数据及执行实体元数据，详见图2。

数据集元数据聚焦描述与数据内容紧密相关的各类信息，涵盖25个元数据元素。这些元素从不同角度对数据本身的内容特性进行规范，为了解数据提供丰富的信息支撑。

活动元数据针对数据溯源过程中涉及的各类活动相关信息展开描述，包含3个元数据元素。通过这些元素记录数据在溯源过程中所历经的各类活动情况，为追溯数据的流转轨迹奠定基础。

执行实体元数据主要围绕数据活动所涉及的人员、工具等相关信息进行描述，包含8个元数据元素。借助这些元素可明确数据活动中参与的主体及所运用的工具等核心要素，有助于精准把握

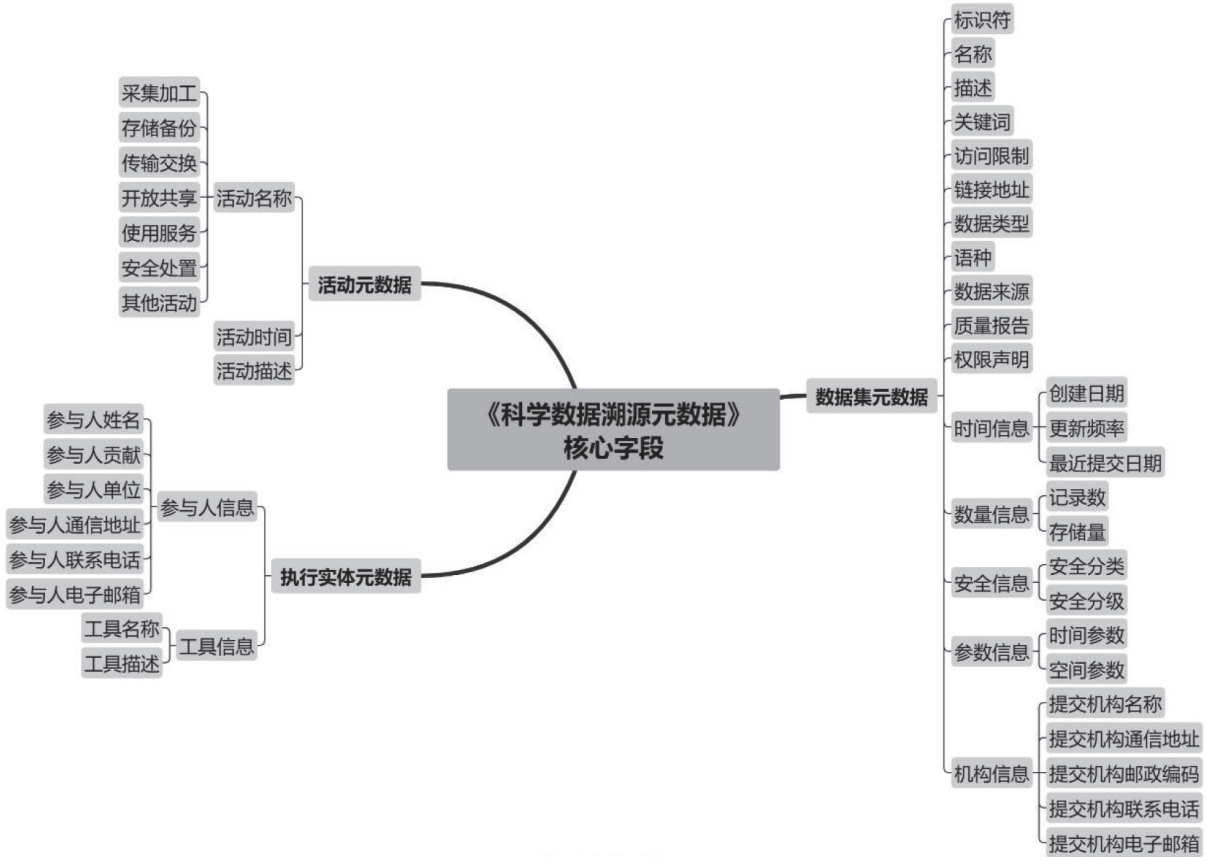

图2《科学数据溯源溯源元数据》核心字段

数据活动中的责任主体与技术依托情况。

4 标准实践与推广

早在2019年，标准起草团队便参与烟草科研大数据资源体系与数据标准体系研究重大专项，研制了烟草科研大数据标准体系中的《数据溯源元数据》项目标准[11]。该项目标准对烟草科研大数据的数据溯源描述模型中3个基本类（即数据、活动和执行实体）的元数据信息予以界定，适用于烟草科研大数据管理系统，为烟草科研大数据平台的建设提供切实有效的指导。

在《科学数据溯源元数据》国家标准公开征求意见阶段，进行多方意见征求。一方面向51个国家科技资源共享服务平台（包含20个国家科学数据中心、31个国家生物种质与实验材料资源库）征求意见；另一方面向全国网络安全标准化技术委员会（SAC/TC 260）进行定向征求意见。通过广泛且有针对性的意见征集，各方在实践层面达成了共识，为该标准后续的完善与推广奠定了良好基础。

2025年1月，《科学数据溯源元数据》国家标准正式发布。中国科学院网站率先对此进行了报道[12]，发挥了重要的信息传播作用。而作为该标准的第一完成单位——中国科学院计算机网络信息中心更是积极履行职责，大力推动标准的应用推广工作，并通过相关公众号及网站平台等渠道进行宣传[12-13]，为其在学科领域的应用创造有利条件。

随后，标准起草团队为拓展应用、深化成果，将该标准的相关成果应用于国家生物信息中心项目研制工作。结合生物数据自身特点与特征，制定《生物数据溯源元数据》项目标准。采用元数据描述元素记录生物数据在整个生存周期内的演变

信息及演变处理内容。为评估生物数据质量及保障数据安全提供一种行之有效的解决方案，显示出该标准成果在不同领域的应用价值。

展望未来，中国科学院计算机网络信息中心将持续推广《科学数据溯源元数据》，充分发挥其在科学数据管理中的基础支撑作用。通过不断扩大标准应用范围，有望改善科学数据质量，切实提升科学数据安全水平，为推动我国科学数据治理工作的高质量发展贡献力量。

5 结 语

GB/T 43707—2025《科学数据溯源元数据》国家标准作为科学数据管理领域的一项重要的基础性标准，围绕科学数据溯源所涉及的数据集元数据、活动元数据及执行实体元数据，构建起一套严谨且系统的规范，为科学数据质量管理实践活动提供了坚实的依据。该标准与 GB/T 34945—2017《信息技术 数据溯源描述模型》存在紧密且重要的联系，是后者在科学数据治理场景的有力补充，适用于科学数据全生存周期中数据安全、数据质量的溯源管理。随着实践应用的不断深入推进，标准仍有进一步完善的空间与必要。例如，积极探索领域扩展方案，让标准能够更好地适配不同领域需求，在更广泛的科学数据管理场景中发挥作用。

参考文献

[1] 苏靖.大数据时代加强科学数据管理的思考与对策[J].中国软科学,2022(9):50–54.

[2] 李国杰.大数据与计算模型[J].大数据,2024,10(1):9–16.

[3] 赵一鸣,谭欣佩,张胜发.科学数据价值评估研究[J].图书情报工作,2025,69(6):58–71.

[4] 信息技术 数据质量评价指标：GB/T 36344—2018[S].

[5] The Open Provenance Model (v1.00) [EB/OL]. [2025–5–22]. https://openprovenance.org/opm/.

[6] SAHOO S S, BARGA R S, GOLDSTEIN J, et al. Provenance algebra and materialized view–based provenance management[C]. In: Proceedings of the 2nd International Provenance and Annotation Workshop. Berlin: Springer, 2008: 531–540.

[7] Olaf Hartig: Provenance Information in the Web of Data [C]. In: Proceedings of the Linked Data on the Web (LDOW) Workshop at WWW. Madrid, Spain. 2009.

[8] 信息技术 数据溯源描述模型：GB/T 34945—2017[S].

[9] 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.

[10] 科学数据溯源元数据：GB/T 43707—2025[S].

[11] 胡良霖,朱艳华,高瑜蔚,等.烟草科研大数据标准体系的构建[J].烟草科技,2020,53(4):100–106.

[12] 《科学数据安全分类分级指南》等5项国家标准发布[EB/OL].(2025–02–17)[2025–05–26]. https://www.cas.cn/yx/202502/t20250213_5047018.shtml.

[13] 中心牵头研制的国家标准《科学数据溯源元数据》正式发布 [EB/OL].(2025–02–20)[2025–05–28]. https://mp.weixin.qq.com/s/1tZ7Fm1WffMUfuvdAA4FEA.