

数字保存中信息对象真实性的保障机制研究*

——以人口健康科学数据长期保存为例

王茜 王蕾 胡佳慧 杨晨柳 姚宽达 方安
(中国医学科学院医学信息研究所, 北京 100020)

摘要: 从信息对象保存的生命周期不间断管理、元数据管理与监测、真实性影响因素管理3个维度对现有信息对象真实性保障模式进行梳理与归纳, 立足可信数字仓储对于数字保存中信息对象真实性保障的基本要求, 基于人口健康科学数据长期保存系统建设实践, 从制度、技术等维度探讨涵盖人口健康科学数据提交、接收和长期保存全链条的真实性保障措施体系。

关键词: 科学数据; 长期保存; 可信数据仓储; 数字对象真实性; 数据监管

中图分类号: G354.49; G255.2 **DOI:** 10.3772/j.issn.1673-2286.2021.04.002

引文格式: 王茜, 王蕾, 胡佳慧, 等. 数字保存中信息对象真实性的保障机制研究——以人口健康科学数据长期保存为例[J]. 数字图书馆论坛, 2021 (4) : 11-16.

1 数字保存中信息对象真实性保障理论认知与现有模式

1.1 数字保存中信息对象真实性内涵界定

关于数字保存中信息对象的真实性, 业界将其界定为一个与信息对象状态、传递模式以及长期保存方式相关的、基于证据判断的问题, 着眼于信息对象来源正确和长期保存过程真实两个维度。如Caplan^[1]将数字保存中信息对象真实性界定为与其最初状态的原始对象相比, 被保存的信息对象是完整且未被非授权修改的一种状态^[2]; CCSDS^[3]、张智雄^[4]、臧国全^[5]、CTS^[6]等提出, 可以从信息对象来源可信、未发生非授权的内容修改、内容完整性得到有效维护、提供信息对象真实性验证机制等维度进行验证。国际标准化组织^[7-8]、德国数字资源长期存储专业网络^[9]、国际科学理事会世界数据系统、中国国家档案局^[10]等主张从数字资源保存周

期角度对信息对象真实性进行界定与分析。

1.2 数字保存中信息对象真实性保障的基本要求

数字保存中信息对象真实性保障基本要求主要分布于数字资源保存组织发布的标准规范之中。因此, 本文重点梳理与研究国际标准化组织发布的《可信数字仓储的审计与认证指标》(ISO16363)、德国数字资源长期存储专业网络发布的《可信数字仓储指标目录》(Nestor)、国际科学理事会世界数据系统发布的《Core Trust Seal可信数据仓储要求(2020—2022年)》和《Core Trust Seal可信数据仓储要求: 扩展指南(2020—2022年)》^[11]、中国国家档案局发布的《文书类电子档案检测一般要求》(DA/T70-2018)^[12]等标准规范关于数字保存中信息对象真实性保障措施体系, 具体如下。

*本研究得到国家科技图书文献中心委托子任务“医学领域规范数据加工和丰富化”资助。

(1) 信息对象接收阶段。数字仓储需要从信息对象和数据提交者两个层面开展真实性管理。一方面,要求数字仓储提取和保留其所要保存信息对象的内容及其信息属性,并提供用于验证完整性和正确性的机制和过程;另一方面,进行数据提交方真实性核实与验证。

(2) 信息对象摄入阶段。要求数字仓储采取提交方在授权机构进行正式注册、上传过程中使用数字签名、保存对象来源元数据真实性评估等管理措施,保障信息对象摄入前、摄入中、摄入后的真实性。

(3) 信息对象保存阶段。要求数字仓储采取建立信息对象真实性审计与回溯机制、记录信息对象(包括元数据)保存阶段所有转换(更改或删除)操作的完整信息等措施。

(4) 信息对象访问阶段。数字仓储应该能够向用户证实自身是用户使用的信息对象的提供者,如使用数字签名技术交付使用用户。在信息对象真实性不明确的情况下,应该向用户提供信息对象的相关资料,如建立记录保存过程中关于信息对象来源和所有更改的元数据体系,以服务于真实性评估。

1.3 数字保存中信息对象真实性保障的现有模式

数字保存中信息对象真实性保障在实践层面探索出以下3种典型模式。

(1) 基于信息对象保存生命周期的不间断管理模式。该模式的基本思路:基于信息对象保存生命周期理论和信息对象连续性管理理论^[13-14],把数字保存中的真实性维护落实到信息对象保存生命周期全过程,在信息对象接收、摄入、保存和服务阶段采取制度、技术、程序及工具等真实性保障措施体系。

(2) 基于信息对象保存元数据的管理与监测模式。该模式的基本思路:面向数字保存中信息对象真实性的基本需求,即向当前和未来的用户保证由存储库保存和分发的信息对象是最初摄入存储库的信息对象的原始副本,或者自摄入后对原始信息对象所进行的任何授权修改都被完整地记录,监测和管理与信息对象真实性相关的元数据以实现信息对象真实性保障目标^[5, 15-17]。

(3) 基于信息对象保存真实性影响因素管理模式。该模式的基本思路:着眼于数字保存中信息对象真实性的影响因素,从环境监测、制度规范、技术系统、

人员素质、管理行为等方面构建涵盖数字保存生命周期的真实性保障措施体系^[14, 18-20]。

2 人口健康科学数据长期保存信息对象真实性保障需求分析

2.1 人口健康科学数据长期保存信息对象的基本特点

人口健康科学数据长期保存的信息对象是由科研项目、科研机构、科研人员产生的生物医学科学数据集合,包括科学数据描述信息、科学数据实体、科学数据辅助工具软件等类型的资源。其中,科学数据描述信息包括科学数据实体的相关描述信息、样例数据、数据字典及伦理、人遗等过程资料;科学数据实体是指科学研究过程中的原始数据和衍生数据所形成的完整数据库或数据集;科学数据辅助工具软件是指用于科学数据处理、加工和分析的专门辅助软件工具等,包括工具软件本身和相关使用说明。相较于期刊文献资源,人口健康科学数据长期保存的科学数据实体对象数据类型多样,数据格式复杂,涉及文本数据、数值数据、图形数据、影像数据等类型^[21],涵盖dmp、mdb、pdf、sav、sql、table、txt、xls等文件格式^[22],存在结构化与非结构化的不同结构化情况,部分数据包体量达100GB及以上;与单一类型的科学数据相比,人口健康科学数据涵盖化学、生物学、心理学、水产学、基础医学、临床医学、药学、工程与技术科学、体育科学、统计学等多个一级学科^[22]。此外,人口健康科学数据资源创建者也较为多样,包括国家部委、医院、高等院校、研究机构等。

2.2 人口健康科学数据长期保存信息对象真实性保障设计思路

人口健康科学数据长期保存信息对象的真实性是指人口健康科学数据长期保存信息对象的内容、结构及背景信息,在经过提交、接收、保存及迁移等过程后,仍与提交时的状态一致。针对人口健康科学数据长期保存信息对象的基本特点,结合长期保存生命周期场景,围绕信息对象提交、接收和保存管理的关键环节,人口健康科学数据长期保存系统从信息对象来源可信、接收内容真实、保存过程完整3个维度,设计人口

健康科学数据长期保存信息对象的真实性保障措施体系(见图1)。

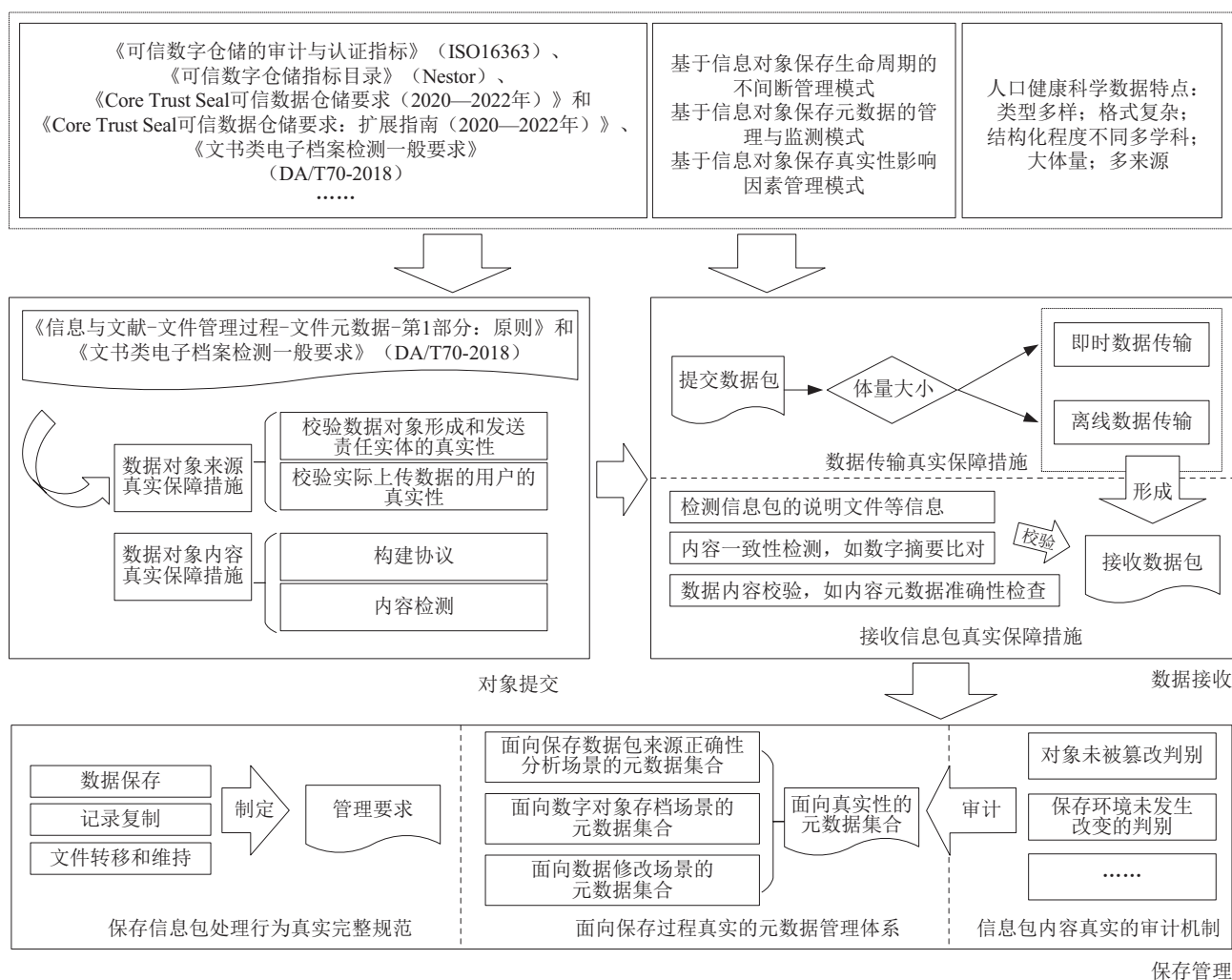


图1 人口健康科学数据长期保存信息对象真实性保障设计思路

3 人口健康科学数据长期保存信息对象真实性保障措施体系

3.1 信息对象来源可信的保障措施

为了应对人口健康科学数据资源创建者多样的挑战, 确保提交到人口健康科学数据长期保存系统中的信息对象真实可靠, 人口健康科学数据长期保存系统主要从信息对象来源真实和信息对象内容真实两个方面制定相关保障措施。

(1) 信息对象来源真实的保障措施。借鉴《信息与文献-文件管理过程-文件元数据-第1部分: 原则》(GB/T 26163.1—2010)中关于文件形成和发送阶段文件真实性的要求^[23], 考虑人口健康科学数据长期保存

科学数据资源创建者多样的特点, 从用户创建、用户身份认证等环节, 开展信息对象形成和发送责任实体(具有法人性质提交机构)、实际上传数据的用户等的真实性校验。

(2) 信息对象内容真实保障措施。信息对象内容真实保障措施包括构建协议和内容检测两个主要手段。一方面, 通过协议明确数据包的描述信息包括但不限于人口健康科学数据集名称、数据集创建者、数据集责任单位、文件大小、文件格式等要素, 用于开展信息对象来源正确性审查, 服务于内容真实性检测。另一方面, 借鉴《文书类电子档案检测一般要求》(DA/T70-2018)关于归档环节电子文件真实性检测基本规范^[24], 要求数据提交方在数据提交之前对信息对象进行来源、内容、信息包3个维度的真实性检测, 并提供检测结

果报告。该检测结果报告是接收、摄入、保存等环节真实性保障的重要依据。此外,针对结构化的人口健康科学数据,内容检测还从内容元数据准确、描述元数据与内容关联一致角度进行深度评价。

3.2 接收内容真实的保障措施

接收内容真实的保障机制:主要通过数据传输和接收信息包的真实保障措施,保证移交和传输到人口健康科学数据长期保存系统中的数字对象的原始性和真实性。具体措施如下。

(1) 数据传输真实保障措施。由于人口健康科学数据分散于不同的机构中,系统接收数据包依赖网络传输。因此,充分考虑网络环境和网络传输协议,形成面向不同体量数据包的即时数据传输方案和离线数据传输方案。GB级别的数据实体应采用由数据离线上、数据离线下、建立数据缓存服务器等方式组成的离线数据传输方案^[25]开展数据传输,从而确保数据传输过程不丢包。同时,采用加密等技术手段,保证科学数据在传输过程中不被篡改。

(2) 接收信息包真实保障措施。接收信息包真实保障措施是从接收包结构、内容、摘要等维度开展接收包的真实性分析,具体包括:对接收的人口健康科学数据信息包的说明文件和目录文件、信息包目录结构等进行规范性检测,以保证移交信息包的组织结构和内容符合移交规范;采用数字摘要比对、数字签名等方式对接收信息包的一致性进行检测,确保待接收数据包与提交数据包相一致并且内容真实性、数据未被篡改;捕获并校验科学数据接收信息包名称、大小、格式、创建时间等内容属性信息。另外,对结构化的科学数据包开展内容元数据准确的校验。

3.3 保存过程完整的保障措施

保存过程完整的保障机制包括制定支撑保存信息包真实完整的处理行为规范、构建面向保存过程真实的元数据体系、设计支撑保存信息包内容真实的审计机制三重手段,保证人口健康科学数据长期保存系统中信息对象的真实和完整。具体措施如下。

(1) 制定支撑保存信息包真实完整的处理行为规范。保存信息包处理行为真实完整规范定义不同保存环节的数据管理要求。信息对象保存环节中,要求监测

保存信息包自创建以来的所有变化信息。保存信息包转移、维护环节中,要求保存信息包的完整性,并实施安全监控。保存信息包备份环节中,要求明确备份日期及负责人,明确从创建者处获得的信息包与保存系统副本之间的关系,分析备份过程对其形式、内容、可及性和使用的影响。此外,在保存系统的副本不能完全、可靠地再现其真实性和完整性时,系统需要记录这些信息并随保存信息包一并保存。以不同保存环节的具体要求为依据,建立保存对象的访问权限,进行保存信息包删除和迁移风险控制,建立防止丢失、损坏及媒体脆弱性、技术过时的相应措施。

(2) 构建面向保存过程真实的元数据体系。保存过程真实的元数据体系是基于保存过程涉及的主要场景而构建的涵盖保存对象、保存时间和保存业务主体描述信息的集合。面向保存信息包来源正确分析场景,设计信息对象描述信息、信息对象签名信息和数字摘要信息3个维度的元数据集合。其中,信息对象描述信息包括人口健康科学数据集的名称、大小、包含的文件数量;信息对象签名信息包括签名编码、签名算法等内容;数字摘要信息包括摘要编码值、编码生成时间、摘要生成、编码算法。面向信息对象存档场景,元数据集合包括保存对象描述信息、存储路径、保存人、保存时间、备份信息、存储介质信息。其中,保存对象描述信息不仅包括人口健康科学数据长期保存信息对象描述信息,还包括人口健康科学数据的文件格式、支持数据查看的软件名称、软件版本、软硬件运行环境;备份信息包括备份地址、备份与存档关联关系、备份创建时间;存储介质信息包括介质类型、位置、名称、生产厂家、生产日期以及存储介质的生命周期。面向保存信息对象修改场景,元数据集合包括修改人、修改时间、修改内容、修改前后的摘要值、修改前后的摘要算法、授权修改依据等内容。

(3) 设计支撑保存信息包内容真实的审计机制。支撑保存信息包内容真实的审计机制是以面向保存过程真实的元数据为对象,开展保存对象未被篡改、保存环境未发生改变的真实性判别,从而尽早探测数据面临的真实性风险。以保存对象未被篡改审计为例,一方面根据协议约定周期,定期检测、比对人口健康科学数据集数据包文件数量、文件大小、数据集摘要的数值,若文件数据量、数据集大小、摘要信息与保存对象记录的元数据不一致或与协议不一致则认为数据对象可能被篡改;另一方面,以数据包记录的备份信息相关元数据为依

据,判断备份数据包的数据量、数据集大小等与保存对象记录的元数据是否一致,并检测保存路径是否正确、保存硬件是否运转等,若备份信息不一致或备份路径不可访问则认为保存对象的备份存在真实性风险。

4 结论与展望

人口健康科学数据作为卫生健康领域关键的战略性基础资源,为医学创新发展、临床诊疗和疾病预防等提供着重要支撑^[26],保障人口健康科学数据的真实可用性成为对其进行管理的核心目标之一,多个数据管理机构均将其纳入工作布局之中。如美国国立医学图书馆在其发布的规划中明确,将通过科研数据存储库开发、政策标准制定和基础设施建设等措施,对大型科研项目(如All of Us、BRAIN Initiative)的科研数据进行存储和管理^[27];哈佛大学图书馆生物医学数据管理服务中将生物医学数据重现性需求作为重要服务内容^[28]。为了支撑人口健康科学数据价值的持续、高质量地发挥,立足人口健康科学数据长期保存系统建设实践,基于信息对象长期保存生命周期,着眼于信息对象长期保存的接收、摄入、保存核心关键流程,从制度、技术等方面对人口健康科学数据长期保存信息对象真实性保障策略进行研究。

未来,面向人口健康科学数据长期保存系统建设实践和工程化需求,需要以制度和技术的为双重抓手,优化、解决人口健康科学数据长期保存信息对象真实性保障的问题。在制度方面对人口健康科学数据长期保存进程中信息对象真实性保障机制和策略进一步深化。如将人口健康科学数据长期保存真实性保障机制向前延伸,深入到人口健康科学数据生命周期管理的前端,积极参与科研主体科学数据管理工作和规范的制定与实施;又如结合人口健康科学数据长期保存系统研发与应用过程,不断调整、完善、具化真实性评价内容。同时,也要探索新技术与人口健康科学数据长期保存信息对象真实性保障机制的结合点,提升真实性保障机制实施的效果和效率。如利用基于区块链技术^[29-30]和理论进一步探究技术层面人口健康科学数据长期保存信息对象真实性保障的技术机制的设计与实现。

参考文献

- [1] CAPLAN P. Understanding PREMIS (2009) [EB/OL]. (2012-01-10) [2021-02-02]. <http://www.loc.gov/standards/premis/understanding-premis.pdf>.
- [2] 赵安琪,付少雄,冯亚飞. 国外健康科学数据管理实践及启示[J]. 图书情报知识, 2020 (1): 105-114.
- [3] CCSDS. The Consultative Committee for Space Data Systems. Audit and Certification of Trustworthy Digital Repositories [EB/OL]. [2020-11-23]. <https://public.ccsds.org/pubs/652x0m1.pdf>.
- [4] 张智雄. 数字资源长期保存技术的研究与实践[M]. 北京: 国家图书馆出版社, 2015.
- [5] 臧国全,崔巍,李哲. 数字保存的真实性风险检测[J]. 情报理论与实践, 2018, 41 (6): 28-33, 38.
- [6] Core Trust Seal Standards and Certification Board. Core Trust Seal Trustworthy Data Repositories Requirements 2020-2022 [EB/OL]. [2020-11-22]. <https://zenodo.org/record/3638211#.X7oj9M0zbic>.
- [7] InterPares. Requirements for Assessing and Maintaining the Authenticity of Electronic Records [EB/OL]. [2020-10-29]. http://www.interpares.org/book/interpares_book_k_app02.pdf.
- [8] ISO 14721: 2012 Space data and information transfer systems- Open archival information system (OAIS) - Reference model [S/OL]. <https://www.iso.org/standard/57284.html>.
- [9] Nestor Working Group. Catalogue of Criteria for Trusted Digital Repositories [EB/OL]. [2020-11-01]. https://files.dnb.de/nestor/materialien/nestor_mat_08-eng.pdf.
- [10] 国家档案局. 科学技术研究档案管理规定 [EB/OL]. [2020-12-10]. <https://www.saac.gov.cn/daj/bmgz/202010/6580c849e8db4417888d9485156e43d1/files/9d48c140690f4aa388bd0f914d755f43.pdf>.
- [11] Core Trust Seal Standards and Certification Board. Core Trust Seal Trustworthy Data Repositories Requirements: Extended Guidance 2020-2022 [EB/OL]. [2020-11-22]. <https://zenodo.org/record/3632533#.X7ozas0zbic>.
- [12] 王大众. 《文书类电子档案检测一般要求》解读[J]. 中国档案, 2019 (4): 34-35.
- [13] 方昀,潘未梅. InterPARES项目对电子文件可靠性真实性保障的哲学基础和理论基础[J]. 档案学研究, 2013 (3): 36-40.
- [14] 张宁,王朝伟,柳增寿,等. 基于数字连续性思想的电子文件真实性评估策略研究[J]. 档案学研究, 2015 (6): 69-72.

- [15] VERMAATEN S, BRIAN L, PRISCILLA C. Identifying Threats to Successful Digital preservation: The SPOT Model for Risk Assessment [EB/OL]. [2020-06-06]. <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>.
- [16] 张正强. 论文件管理元数据结构对保证电子文件真实性的标准化作用与意义 [J]. 档案建设, 2011 (11): 4-7.
- [17] 骆建珍, 杨安荣, 马来娣. 电子档案“四性”检测要求及其实现方法 [J]. 浙江档案, 2017 (12): 27-30.
- [18] 孙刚. 电子文件真实性保障模型研究 [D]. 天津: 天津师范大学, 2007.
- [19] 周天舒, 李劲松, 杨一兵, 等. 区域医疗系统数据真实性保障流程优化 [J]. 浙江大学学报 (工学版), 2010, 44 (8): 1484-1489, 1507.
- [20] 石进, 薛四新, 赵小柯. 基于区块链技术的电子文件真实性保障系统模型研究 [J]. 图书情报知识, 2019 (6): 111-119.
- [21] 吴思竹, 钱庆, 周伟, 等. 面向人口健康领域科研项目数据汇交的数据仓储设计与实现 [J]. 数据分析与知识发现, 2020 (12): 2-12.
- [22] 尹岭, 付磊. 人口健康领域科技基础性工作数据资料集成与规范化整编 [M]. 北京: 科学出版社, 2020: 78.
- [23] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. 《信息与文献-文件管理过程-文件元数据-第1部分: 原则》(GB/T 26163.1-2010) [S]. 北京: 中国标准出版社, 2011.
- [24] 中华人民共和国国家档案局. 文书类电子档案检测一般要求 (DA/T70-2018) [EB/OL]. [2020-06-06]. <https://www.saac.gov.cn/daj/hybz/201806/51dc58c8f5864df49eaacb82dfe8e193/files/c56f09e7648345baa692f82b4bdc4689.pdf>.
- [25] 涂平, 朱晓铃, 满旺. 大数据量数据网络传输研究与实现 [J]. 微型机与应用, 2010, 29 (7): 6-9.
- [26] 关键. 医学科学数据共享与使用的伦理要求和管理规范 (九): 医学科学大数据及其属性 [J]. 中国医学伦理学, 2020, 33 (12): 1427-1431.
- [27] 赵栋祥. 《美国国家医学图书馆战略规划: 2017—2027》的解读与启示 [J]. 图书情报工作, 2019, 63 (5): 138-146.
- [28] 叶冬梅, 牛卫东. 国外高校图书馆医学科学数据管理服务调研与启示 [J]. 图书馆学研究, 2020 (5): 67-74.
- [29] 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.
- [30] 黄新平. 基于区块链的政府网站信息资源安全保存技术策略研究 [J]. 图书馆, 2019 (12): 1-6.

作者简介

王茜, 女, 1981年生, 博士, 副研究馆员, 研究方向: 信息技术应用。

王蕾, 女, 1989年生, 硕士, 助理研究员, 研究方向: 信息技术、大数据处理。

胡佳慧, 女, 1987年生, 博士, 副研究员, 研究方向: 医学科研数据管理与服务研究、医学数据分析与知识发现。

杨晨柳, 女, 1991年生, 硕士, 助理研究员, 研究方向: 医学信息组织与管理、医学数字资源长期保存。

姚宽达, 男, 1993年, 硕士, 研究实习员, 研究方向: 医学信息组织与管理、医学数字资源长期保存。

方安, 男, 1976年生, 博士, 研究馆员, 通信作者, 研究方向: 医学知识组织与数字图书馆, E-mail: fang.an@imicams.ac.cn.

Research on the Authenticity Guarantee Mechanism of Information Objects in Digital Preservation: Take the Long-Term Preservation of Population Health Science Data as an Example

WANG Qian WANG Lei HU JiaHui YANG ChenLiu YAO KuanDa FANG An
(Institute of Medical Information/ Medical Library, CAMS & PUMC, Beijing 100020, China)

Abstract: Sorts out and summarizes the existing digital object authenticity guarantee mode, and faces credible numbers from the uninterrupted management of information object preservation life cycle, information object preservation metadata management and monitoring, and information object preservation authenticity influencing factor management. Faces to the basic requirements for authenticity assurance of digital preservation of information objects for credible digital storage and bases on the construction practice of the long-term preservation system of population health science data, from the system and technology, the authenticity guarantee system covering the entire chain of long-term preservation, submission, reception and preservation of population health scientific data was discussed.

Keywords: Scientific Data; Long-Term Preservation; Trust Worthy Digital Repositories; Digital Object Authenticity; Data Curation

(收稿日期: 2021-03-12)