• 医学信息组织与利用 •

医学数字资源长期保存存储策略研究*

胡佳慧 方 安 杨晨柳 范云满 高 星

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 从政策法规、战略储备、成果验证以及资源利用 4 个方面，分析医学数字资源长期保存的必要性。基于 OAIS 参考模型，对存储功能实体及实体间的相互关系进行梳理。通过对存储模式、存储架构、存储技术和保障机制的研究，探索科学的存储策略，以实现医学数字资源的长期可靠存储。

〔关键词〕 存储策略；长期保存；医学数字资源

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2017.12.016

Study on the Long－term Preservation Storage Strategy of Medical Digital Resources *HU Jia－hui, FANG An, YANG Chen－liu, FAN Yun－man, GAO Xing, Institute of Medical Information of Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The necessity of long－term preservation of medical digital resources is analyzed from 4 aspects in the paper, namely, policies and regulations, strategic reserves, scientific findings verification and resource utilization. Based on the OAIS reference model, the storage function entities and the relationships are clarified. Through study of storage mode, architecture, technology and guarantee mechanism, scientific storage strategy is explored to realize the long－term reliable storage for medical digital resources.

〔**Keywords**〕 Storage strategy; Long－term preservation; Medical digital resource

1 引言

数字化为医学信息资源的管理、利用与交互带来了极大的便利。然而数字资源的生存对其保存技术和保存环境有着较强的依赖性，容易受到外界的影响而导致数据的破坏和信息的丢失。档案馆、博物馆和图书馆率先意识到开展数字化资源长期保存的重要性。2013 年 12 月联合国教育、科学及文化组织（United Nations Educational, Scientific and Cultural Organization, UNESCO）、国际图书馆协会联合会（International Federation of Library Associations and Institutions, IFLA）与国际档案理事会（International Council on Archives, ICA）联合发起 PERSIST（Platform to Enhance the Sustainability of the Information Society Transglobally）项目[1]，旨在为全球文化遗产提供长期访问和可信赖的保护。

在国家科技图书文献中心（National Science and Technology Library, NSTL）的推动下，代表我国图书馆届加强文献资源在本土长期保存愿望的《数字文献资源长期保存共同声明》于 2015 年 9 月发布[2]。围绕国家数字科技文献资源长期保存体系建设，目前已建成中国科学院文献情报中心、中国科学技术信息研究所和北京大学图书馆 3 个保存节

〔修回日期〕 2017－09－26

〔作者简介〕 胡佳慧，助理研究员，博士；通讯作者：方安，副研究馆员。

〔基金项目〕 中国医学科学院中央级公益性科研院所基本科研业务费项目“医学数字资源长期保存策略研究”（项目编号：2016ZX330022）。

点。长期保存的研究和实践在档案和图书文献领域已趋于成熟。然而相较于资源的生产速度，资源的保存能力远远滞后。在大数据时代占据较大构成比例的数据资源（如科学数据等）的长期保存策略尚待研究和探索。

医学数据种类繁多，包括各种临床数据、组学数据、环境数据、疾病数据、人口统计学数据等[3]。随着大数据计算、分析与挖掘技术的发展，医疗卫生服务需求日益增长，医学数据成为支撑数据密集型科学发现的重要资源，为基于医学大数据的临床预测、检验、诊疗以及流行病学监测等应用[4]提供研究基础。医学领域对应用环境和服务有着特殊的需求，例如医疗事故、经济纠纷、隐私维权等的追踪和问责，要求对诊疗过程有完整的记录并得到妥善长久保存。基于医学领域资源建设和信息服务的发展需求，有必要开展医学数字资源的长期保存，以增强医学数字资源战略保障及服务能力。

数字资源长期保存的目标是实现保存资源的真实性、可靠性以及长期可解释性。开放归档信息系统（Open Archival Information System，OAIS）[5]提供了通用的长期保存参考框架。在 OAIS 参考模型中，存储作为数字资源保存的重要环节，是长期保存的核心功能实体之一。长期保存要求在实现资源存储的同时，对保存内容和保存环境进行主动监管，在存储策略的科学指导下，维护数字内容的不变性，应对长期保存过程中环境、技术等的变化。为确保医学数字资源的长期利用价值，本研究从政策法规、战略储备、成果验证以及资源利用 4 个方面，剖析医学数字资源长期保存的必要性。鉴于存储在长期保存过程中的关键性地位，基于业界广泛认可的 OAIS 参考模型，梳理存储功能及各实体间的相互关系。通过对存储模式、存储架构、存储技术和保障机制的研究，探索适合医学领域的资源长期存储策略。

2 医学数字资源长期保存必要性分析

2.1 国家政策法规的明确要求

基于医学领域特殊的应用环境和服务需求，医学数据的保存受到国家层面的重视。国家卫生计生委先后于 2016 年 8 月和 2017 年 2 月分别印发《医学影像诊断中心管理规范（试行）》[6]和《电子病历应用管理规范（试行）》[7]，明确影像资料的保存时间为 10 年以上，且至少保证 3 年在线，门（急）诊电子病历的保存时间分别不少于 15 年，住院电子病历的保存时间不少于 30 年。随着医疗卫生事业的发展，医学数字资源将在日趋健全的规章制度下得到长期有效保存。

2.2 医学资源战略储备的重要手段

由于数字资源自身的脆弱性，且易受环境的影响，造成数据的损坏和信息的丢失[8]。在长期的医学研究、探索和实践过程中，医学数字资源具有切实的长期保存需求[9]。2017 年 5 月“蠕虫式”勒索病毒软件 WannaCry 影响全球范围内的数据安全，波及包括医疗在内的多个重要行业。长期保存是医学资源战略储备的重要手段，通过开展科学有序的主动保存活动，确保数字资源的持久生存能力，为重要医学资源在必要条件下的可靠获取创造条件。

2.3 医学成果可验证性的有力依据

医学领域注重知识的积累和验证[10]，对资源的真实性和准确性有着较高的要求。长期保存不仅要求对数字内容本身的保存，还要求保证保存生命周期过程中资源的真实性和完整性。OAIS 数字保存框架提供了基于统计和内容的数字对象不变性检查方法，起源信息负责记录对原始保存数字对象的各项操作，数字资源的可追溯性具有保障。此外，ISO 16363 标准[11]从组织机制、数字对象管理以及基础设施等方面为可信赖的数字仓储认证提供准绳。

2.4 医学资源长效利用的根本保障

资源保存的根本目的是为资源的长效利用提供保障。在医学数字资源的长期利用过程中，技术、媒体和数据格式的变革，以及特定用户团体的变化等都可能对保存数字对象产生影响。OAIS 明确了长期保存是一种对保存对象进行长期维护的行为，其宗旨是确保特定的用户团体正确理解当前保存的数

字对象。以数据驱动为主的新态势下，开展医学数字资源长期保存活动，将促进资源在医学科研、医疗诊断、卫生决策等领域的长效利用。

3 OAIS 存储功能实体

3.1 存档信息包

信息包是长期保存数据对象可理解性的重要保证。OAIS 将长期保存系统中所存储的信息包称之为存档信息包（Archival Information Package，AIP），AIP 是 OAIS 存储功能模块与摄入、访问模块之间交互的基本对象。根据 OAIS 对信息包的描述，AIP 由内容信息、保存描述信息、封装信息以及包描述信息组成，见图 1。内容信息包括数据对象和呈现信息；保存描述信息包括指引信息、起源信息、情景信息、不变性信息和访问权限信息；内容信息与保存描述信息需要封装在一起，通过封装信息对其识别；包描述信息用于对 AIP 的描述，为 AIP 的检索提供便利。

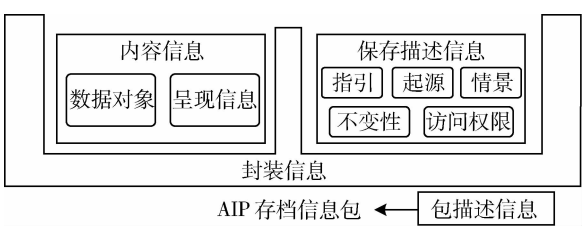

图 1 AIP 结构

3.2 主要功能

3.2.1 数据接收 接收存储请求以及来自摄入功能实体的 AIP。根据存储请求中对 AIP 数据对象使用频率的说明，选择正确的存储媒介。接收任务完成后，向摄入功能实体反馈包含有该 AIP 存储标识的存储确认消息。

3.2.2 存储体系管理 根据存储管理策略、操作统计以及存储请求，通过相应的命令将 AIP 的内容安置在正确的媒介上。遵从 AIP 的各项服务及安全需求，确保 AIP 得到合适的保护。监测错误日志，以确保 AIP 在传送过程中未被损坏。向行政管理功能实体提供对当前存储媒介、可用存储容量以及利用率等的统计报告。

3.2.3 媒介更替 为长期保存提供 AIP 的重生（Reproduce）能力。在媒介更替过程中，不允许改变内容信息和保存描述信息，可以改变组成包信息的数据，但不能造成信息的丢失。迁移策略在选择存储媒介时，必须综合考虑不同类型媒介的预期和实际错误率、性能以及使用成本。

3.2.4 错误检查 基于统计的方法，提供 AIP 完整性保证。要求存储模块中的所有软硬件提供潜在错误通知，这些错误将写入标准的错误日志，由人工检查。PDI 不变性信息提供传输和获取 AIP 过程中内容信息未被更改的保证。此外，还需要追踪和验证所有数据对象有效性的标准机制，以确保数据对象的完整性。

3.2.5 灾难恢复 提供复制数字内容的机制，并选择合适的方式存储副本（如在物理上分离的设备中存储副本）。通过可移动的存储媒介或网络数据传输实现该功能。具体的灾难恢复政策由行政管理功能实体制定。

3.2.6 数据提供 接收 AIP 访问请求，并根据请求中的 AIP 标识以及所要求的媒介类型，提供该 AIP 副本，或将副本传送至一个临时的存储区域。完成后，向访问功能实体发送数据传输通知。

3.3 各功能实体间关系

图 2 从横向和纵向两个维度对存储功能实体进行诠释，包括存储功能实体与 OAIS 其他功能实体（摄入、访问、行政管理）之间的关系，及其内部各功能实体之间的关系。横向提供了 AIP 的数据流向，即存储模块从摄入模块接收 AIP 进行存储，访问模块从存储模块获取所需的 AIP。纵向的存储体系管理、媒介更替、错误检查和灾难恢复功能则为永久存储提供保障。

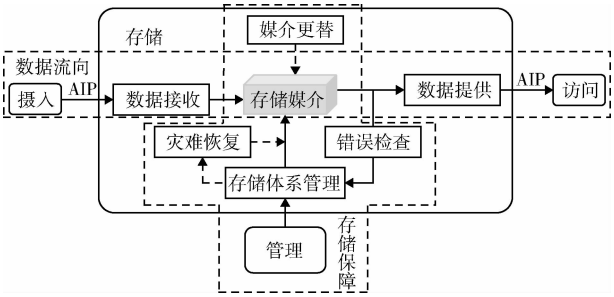

图 2 存储功能实体

4 存储策略

4.1 存储模式

有效的存储策略有赖于对数据集的识别和理解[12]。医学数字资源种类丰富，类型复杂多样，各类医学资源在使用频率、保存成本、保存生命周期等方面表现出不同的特性[13]。对医学数字资源进行长期保存，首先需明确保存资源的特征属性及其对长期保存系统的要求。存储功能模块接收来自摄入模块的 AIP，基于 AIP 内容信息、保存描述信息、封装信息以及包描述信息，对保存对象进行合理分类，在此基础上，选择合适的存储媒介和存储模式，进行永久存储。分级存储模式，见表 1。此外，作为保存生命周期的重要环节，存储阶段需进行主动监管[14]，针对环境的变化、技术的变革、格式的变迁等，及时调整保存策略，确保保存数字内容在长期保存过程中得到妥善存储。

表 1 存储模式

存储模式	代表性媒介	优势	劣势	应用场合	应用示例
在线存储	磁盘阵列	响应速度快	成本代价高	实时性要求较高的应用	医院实时业务数据
近线存储	磁带库	单位容量成本低	存储容量有限	数据周期性更新和访问的应用	阶段性医学成果
离线存储	磁带	寿命长、容量大、成本低	响应速度慢	数据容量要求大、使用频率低的应用	医学灾备数据

4.2 存储架构

长期保存需确保资源在存储过程中的不变性，数据存储的目的是实现资源的永久存储，为资源的长期获取和利用提供真实性和可靠性保证。OAIS 对长期的解释是一段足够长的时间，可以扩展到无限期。因此存储架构的设计需要基于长期保存规划的指导，明确保存资源的存储环境类型（如数据中心、数据仓库、云存储）。针对医学数字资源，数据量呈爆炸式增长，长期保存要求存储架构具备良好的可扩展性、高可靠性以及安全性保证。鉴于单个机构在开展大规模医学数字资源长期保存实践活动中存在的局限性，合作保存通过多个保存机构之间的资源共建共享，可有效缓解技术实现复杂度和昂贵的经济成本所带来的压力，实现资源的高效保存。从保存系统的长期管理和维护角度看，系统自动化可有效降低复杂度，节省人力成本，这要求存储架构需具备一定的自管理和自愈能力。相比传统的集中式存储，分布式存储架构具有良好的可扩展性，且支持协同服务，适合于医学大数据时代资源的长期保存。针对大文件的存储和管理，Hadoop 在吞吐量、可靠性、成本代价等方面表现出优越的性能。分布式文件系统（Hadoop Distributed File System，HDFS）[15]是 Hadoop 的核心元素，采用主/从（Master/Slave）架构实现对集群中所有文件的可靠存储。

4.3 存储技术

长期保存面临着对各种不同类型物理存储资源的整合和利用。软件定义存储突破了传统的在硬件中固化存储功能的局限，以智能化、灵活的方式为面向异构存储资源的长期保存提供了途径[16]。云存储基于按需服务的理念，通过对资源的统一管理与动态调度，实现对存储资源的合理分配和优化处理，通过网络提供按需供给的自动化存储服务。鉴于数字对象保存者对公有云在数据隐私性和服务安

全性方面的担忧[17]，私有云受到了长期保存相关研究机构的关注。欧盟第 7 框架资助的 ENSURE（Enabling Knowledge Sustainability, Usability and Recovery for Economic Value）项目[18]，通过构建面向大规模数据的云存储模型，实现了数字病理切片图像的长期保存。虚拟化是实现云存储的一项重要技术，已在医疗卫生领域得到认可和应用[19]。通过将物理的硬件存储资源虚拟化为逻辑上相隔离的多个资源，虚拟化技术屏蔽了底层硬件的差异，突破了地理空间和数据规模给数据交互和资源共享带来的局限[20]，具有良好的可扩展性和安全性，促成多个机构间的协作保存，并为长期保存过程中的数据迁移以及不同存储策略间的平稳过渡提供便捷性和可操作性。除虚拟化技术外，Docker 容器的持久存储模式[21]给基于云存储的长期保存带来新的机遇。自包含的信息保留格式（Self - contained Information Retention Format, SIRF）[22]为长期保存提供了一种逻辑存储格式，在一定程度上保证了数字对象的不变性[23]。新技术的发展有助于提升长期保存的水平。

4.4 保障机制

长期保存不仅需要实现对资源的保存，还需要确保数字对象在其保存生命周期内的真实性和完整性。安全的基础设施环境为资源的可靠存储提供了基本保障。除最基本的病毒防范措施外，空间数据系统咨询委员会（Consultative Committee for Space Data Systems, CCSDS）[24]补充物理访问、备份、完整性检查方面的安全保障措施，建议利用电子签名等防篡改技术，确保数字内容的可验证性。由于医学资源具有较高的隐私性要求，需根据国家相关法律法规加强保护和处理保存内容中敏感信息的能力和机制。例如远程医疗中的医学影像存储，要求通过数字水印和电子签名等信息安全技术，为版权信息和用户身份提供认证[25]。为实现数据的长期可靠存储，多级多地战略储备库的建立是一项重要的基础性保障策略，以便在极端情况下实现对保存内容的恢复与获取，以灾难预警和应急等方式为保存资源提供安全性保证。为在更长的时间内保持对重要医学记录的持续获取，Mayo 诊所制定了完善的保存计划、迁移策略以及安全保护机制[26]，建立分层存储管理体系，以存储多个备份的方式，为重要资源的长期保存提供保障。此外备份还是长期保存的一种重要技术手段。LOCKSS（Lots of Copies Keeps Stuff Safe）保存系统[27]基于多副本的保存模式，采用 P2P 轮询机制，以低成本的方式实现对保存内容的一致性检查与异常修复，为长期保存提供可靠的安全性保证。

5 结语

医学数字资源具有长期保存的需求，受到国家层面的重视。长期保存是资源战略储备的重要手段，是成果可验证性的有力依据，是资源长效利用的根本保障。根据 OAIS 参考模型，存储模块接收来自摄入模块的 AIP，确保 AIP 在存储阶段的真实性和可靠性，提供给访问模块，是长期保存的核心功能实体。基于对存储策略的研究，医学数字资源种类繁多、规模庞大、隐私性高，需依据切实的保存需求，对资源合理归类并进行分级存储，选择可扩展的分布式存储架构，在新技术的推动下提升长期保存水平，健全安全保障机制，实现医学数字资源的科学有效保存，为医疗卫生事业的长足发展提供可靠资源。

参考文献

1 UNESCO. Digital Sustainability [EB/OL]. [2017 - 08 - 06]. https://www.unesco.nl/digital - sustainability.

2 科技部. 数字文献资源长期保存共同声明发布 [EB/OL]. [2017 - 08 - 06]. http://www.most.gov.cn/kjbgz/201509/t20150928_121823.htm.

3 弓孟春, 陆亮. 医学大数据研究进展及应用前景 [J]. 医学信息学杂志, 2016, 37 (2): 9 - 15.

4 王雪梅, 刘莉. 医学大数据应用前景与挑战 [J]. 医学信息学杂志, 2016, 37 (8): 56 - 60.

5 CCSDS Secretariat. CCSDS 650. 0 - M - 2. Reference Model for an Open Archival Information System (OAIS) [S]. Washington: CCSDS, 2012.

6 国家卫生计生委医政医管局. 国家卫生计生委关于印发医学影像诊断中心基本标准和管理规范（试行）的通知 [EB/

OL]. [2017 - 08 - 06]. http://www.nhfpc.gov.cn/yzygj/s3593g/201608/6622dba2c35f4c88ac05c09e29f877f.shtml.

7 国家卫生计生委医政医管局. 关于印发电子病历应用管理规范（试行）的通知 [EB/OL]. [2017 - 08 - 06]. http://www.nhfpc.gov.cn/yzygj/s3593/201702/22bb2525318f496f846e8566754876a1.shtml.

8 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.

9 胡佳慧, 钱庆, 杨晨柳. 医学数字资源长期保存体系研究 [J]. 医学信息学杂志, 2016, 37 (6): 67 - 73.

10 宁康, 陈挺. 生物医学大数据的现状与展望 [J]. 科学通报, 2015, 60 (Z1): 534 - 546.

11 ISO. Space Data and Information Transfer Systems - Audit and Certification of Trustworthy Digital Repositories [EB/OL]. [2017 - 08 - 06]. https://www.iso.org/standard/56510.html.

12 Apurva Vaidya, Anshul Kosarwal. Big Data Storge [EB/OL]. [2017 - 08 - 06]. https://www.snia.org/events/storage - developer/presentations14.

13 吴艳艳, 唐源, 李霞. 医院 PACS 云存储系统建设 [J]. 医院管理论坛, 2014, 31 (11): 59 - 61.

14 张梦霞, 顾立平. 数据监管的政策研究综述 [J]. 现代图书情报技术, 2016, 32 (1): 3 - 10.

15 The Apache Software Foundation. Apache Hadoop [EB/OL]. [2017 - 08 - 06]. http://hadoop.apache.org/.

16 董晓莉. 软件定义存储：下一代存储在数字资源长期保存中的应用 [J]. 现代情报, 2017, 37 (2): 38 - 43.

17 高建秀, 吴振新, 孙硕. 云存储在数字资源长期保存中的应用探讨 [J]. 现代图书情报技术, 2010, 26 (6): 1 - 6.

18 Braud M, Edelstein O, Rauch J, et al. ENSURE: Long term digital preservation of Health Care, Clinical Trial and Financial data [C]. Lisbon, Portugal: International Conference on Preservation of Digital Objects. 2013.

19 王春容, 曾宇平. 医院虚拟化云平台构建研究 [J]. 医学信息学杂志, 2016, 37 (5): 24 - 27.

20 孙磊, 胡学龙, 张晓斌, 等. 生物医学大数据处理的云计算解决方案 [J]. 电子测量与仪器学报, 2014, 28 (11): 1190 - 1197.

21 Kvaes. DOCKER: STORAGE PATTERNS FOR PERSISTENCE [EB/OL]. [2017 - 08 - 06]. https://kvaes.wordpress.com/2016/02/11/docker - storage - patterns - for - persistence/.

22 SNIA. Self - contained Information Retention Format (SIRF) [EB/OL]. [2017 - 08 - 06]. https://www.snia.org/tech_activities/standards/curr_standards/sirf.

23 董晓莉. SIRF 与长期保存数字对象的不变性研究 [J]. 图书馆杂志, 2017, 36 (3): 69 - 76.

24 CCSDS. Consultative Committee for Space Data Systems [EB/OL]. [2017 - 08 - 06]. http://public.ccsds.org/default.aspx.

25 曾旭, 司马宇. 基于数字水印和数字签名技术的医学影像存储与传输系统安全机制研究 [J]. 医学信息学杂志, 2016, 37 (7): 44 - 46.

26 Gordon E. J. Hoke. Future Watch: strategies for long - term preservation of electronic records [EB/OL]. [2017 - 08 - 06]. http://content.arma.org/IMM/May - June2012/futurewatchstrategiesforlongtermpreservation.aspx.

27 Reich V, Rosenthal D S H. LOCKSS (lots of copies keep stuff safe) [J]. New Review of Academic Librarianship, 2000, 6 (1): 155 - 161.