• 医学信息组织与利用 •

医学数字资源长期保存系统监控策略研究*

钟 明 胡佳慧 吴思竹

(中国医学科学院医学信息研究所 北京 100020)

〔摘要〕 分析医学数字资源长期保存领域的相关标准和研究现状，归纳医学数字对象变化的影响因素，提出医学数字资源长期保存系统的监控策略，主要包括监控数据摄入过程、存档过程、访问过程及系统运行环境4 个方面。

〔关键词〕 长期保存；监控策略；医学数字资源；可信赖

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2018.09.016

Study on the Monitoring and Controlling Strategy of the Long-Term Storage of Medical Digital Resources *ZHONG Ming, HU Jiahui, WU Sizhu, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper analyzes the related standards and current research situation of long-term storage of medical digital resources, summarizes the impact factors of changes in medical digital objects, and sets forward the monitoring and controlling strategy of the long-term storage of medical digital resources, mainly including the four aspects of monitoring data intake process, saving process and access process as well as system operation environment.

〔**Keywords**〕 long-term preservation; monitoring strategy; medical digital resources; trusted

1 引言

随着国家对信息惠民、大数据战略、健康中国2030 等做出一系列部署以及现代信息技术的迅猛发展，医学大数据的应用发展得到快速推进，各类医学数字资源暴发式增长，长期保存的需求越来越迫切。相关机构积极开展医学数字资源长期保存的研究和实践，以促进数字资源的共享和持续利用。然而在医学数字资源长期保存过程中，环境、技术、安全等多种因素的影响会导致数字对象发生变化，从而破坏保存对象的真实性、完整性、可获得性和长期可解释性。因此长期保存活动中需要对保存系统进行严格的全生命周期监控管理，以构建数字对象长期可信赖的保存环境，确保数字对象的重要属性不发生改变或在长期保存政策允许的范围内改变[1]。本文在对数字保存生命周期和长期保存监控实践的研究基础上分析影响医学数字资源发生改变的各种因素，探索可信赖的动态监控管理机制，以期为医学数字资源长期保存实践提供指导。

〔修回日期〕 2018-05-15

〔作者简介〕 钟明，助理研究员，发表论文3 篇。

〔基金项目〕 中国医学科学院医学信息研究所中央级公益性科研院所基本科研业务费专项“医学数字资源长期保存策略研究”(项目编号：15R01 10)。

2 相关研究现状及数字对象变化影响因素

2.1 长期保存监控管理相关模型和标准

开放档案信息系统（Open Archival Information System，OAIS）参考模型对数字资源长期保存系统的信息处理模式和流程、系统功能结构和外围技术环境等进行规范，其功能模型，见图 1[2]。OAIS 定义了长期保存规划、摄入、归档存储、数据管理、访问和系统管理 6 个功能实体模块，并将信息包分为提交信息包（Submission Information Package，SIP）、存档信息包（Archival Information Package，AIP）和分发信息包（Dissemination Information Package，DIP）3 种类型。英国数字保存中心（Digital Curation Centre，DCC）[3]、英国数据仓储（UK Data Archive，UKDA）[4]、美国 DataONE[5]等组织提出数字保存生命周期模型，其核心环节，见表 1。各数字保存生命周期模型都包含 OAIS 参考模型的主要功能，强调对全生命周期的管理来维护数字对象的真实性、完整性和可用性。ISO 16363 是首个可信赖数字仓储审计与认证的国际标准，第 1 级指标包括组织机制、数字对象管理和基础设施 3 个方面[6]。在数字对象管理部分，ISO 16363 按照 OAIS 的主要功能模块，依次从摄入、保存、访问进行分析，要求对 SIP、AIP、DIP 都做监测和校验；在基础设施部分，标准强调对所有数字对象备份的数量和位置进行管理，对与数据、系统、人员和硬件设备有关的安全风险因素保持系统性分析。

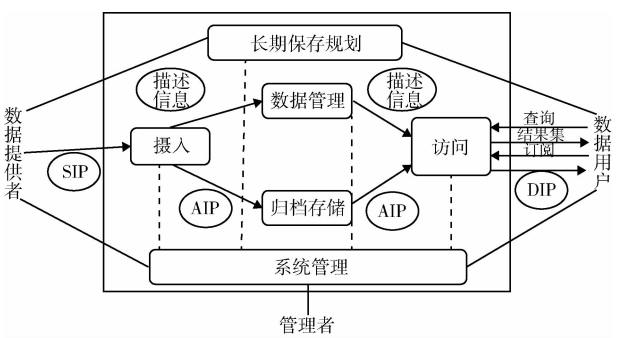

图 1 OAIS 功能模型

表 1 数字保存生命周期模型核心环节

DCC	UKDA	DataONE
概念化	数据创建	数据管理计划
创建或接收	数据处理	数据采集
评估和选择	数据分析	数据质量控制
摄入	数据保存	数据描述
保存活动	数据访问	数据保存
存储	数据重用	数据发现
访问、利用和重用	–	数据整合
转换	–	数据分析

2.2 现有长期保存系统的监控策略

LOCKSS 系统[7]建立在大量副本保证原件的原则上，采用对等网络的轮询和投票机制，要求存档相同内容的多个结点定期计算各自指定内容的消息摘要并进行比较和监控。Fedora 系统[8]采用内容版本管理功能来记录数字对象的变化，即为每次修改的数据流创建 1 个新版本。Fedora 保存数据流的所有版本，提供审计数据流来记录数字对象的所有修改操作。Fedora 可以为每个数据流的每个版本生成 MD5，以便实现数字对象的不变性校验。SCAPE 项目开发的保存监测系统 Scout[9]提供本体知识库来集中管理系统中的保存风险和规避风险所需的信息，采用插件方式集成新的信息源。系统通过浏览知识库、安装触发器来自动通知用户新的风险和机遇，实现对内容配置文件、保存控制政策、PRONOM 注册表和 Web 自动保存 4 个方面的监控。加拿大图书档案馆（Library and Archives Canada，LAC）的可信数字仓储（Trusted Digital Repository，TDR）[10]对摄入区的数字资源进行技术和内容验证。验证通过后 LAC 对档案文件和元数据分别根据分级存储管理和副本策略进行 TDR 存储，对这两个数字仓储执行持续完整性监控、风险格式监控和风险硬件监控。中科院数字资源长期保存系统（Digital Preservation System，DPS）中摄入存档子系统对接收数据进行病毒检查、压缩包可用性检查、MD5 检查并生成新的校验码，对 SIP 的完整性、内容、数量和格式进行检查，利用 Fedora 的版本管理功能对

AIP 进行完整性监控和不变性检查[11]；公共服务子系统提供用户使用统计和恶意下载监控功能，实时监控用户访问行为并做记录。另外 DPS 对存档数据建立多层面、多备份、异质、异地的备份机制。

2.3 数字对象变化影响因素

医学数字资源长期保存活动中导致数字对象变化的因素主要有系统环境、数据处理技术和安全管理因素。系统环境包括系统运行的软件、硬件和外部环境。系统软件、支撑软件和应用软件的可靠性和健壮性，存储载体被干扰、服务器故障和过时、网络通讯故障等风险，系统外部的技术发展、目标群体的改变以及机房温度、湿度的变化都会影响数字对象的真实性和可访问性。数字资源长期保存的处理过程包括数据检查、校验、转换、规范、摄入、保存、分发等多个方面，随着环境变化需要对保存系统进行数据格式迁移、软件环境迁移、仿真等技术处理。这些技术手段在操作过程中有导致数字对象损失的风险。医学数字资源类型多样、规模庞大、层次性和专业性强使保存系统对处理技术有更高要求。数字资源在处理、存储、传输和使用中很容易受到外界干扰、滥用、遗漏和丢失，甚至被非法窃取、篡改和破坏。因此安全管理也是影响数字对象变化的一大因素。医学数字资源包含大量的重要数据和隐私信息，这要求医学数字资源保存系统应具有比一般保存系统更高的安全性保证[12]。

3 医学数字资源长期保存系统监控策略

3.1 摄入过程

3.1.1 数据接收 系统对来源服务器数据进行动态监测，若发现新数据则发送提醒以通知管理员执行原始数据下载。系统监控接收行为，记录接收行为的相关信息，对接收的提交数据按批次进行登记管理，对原始数据按照预定的备份规则实施备份。

3.1.2 数据评估和检查 管理员应对接收的数据进行相关检测、评估和监控，具体包括：对下载的原始数据进行解密、病毒检测、解压缩测试，检查通过后生成 MD5 并通知相关人员；评估数据的效用、时间、形成者和可获取性，确定是否对其进行长期保存；清点各种统计性数据，包括计算各种文档数量、查看文件大小、检验文档路径和检查文档关系，与提供者随包提交的清单进行比较；利用提供方提供的不变性信息对接收的每个数据包或文档进行不变性和完整性检查；验证文件的格式，对于不满足长期保存要求的格式，应在保存原有文件的基础上做必要的格式转换。其中数据的评估过程包括：以社会效用为衡量指标，考察数字资源的内容在利用范围上的广度和在利用价值上的张力；考察数字资源产生的时间，往往新颖的或历史悠久的稀缺资源值得保存；考察数字资源的形成者是否权威、是否经过合法授权，确保数字资源来源可靠；检查存储载体的兼容性、数字资源内容的可识别性，以及数字资源获取的权限是否开放。

3.1.3 数字对象摄入 数据摄入是指将 SIP 转换为 AIP 的过程，该过程需要监控以下 3 个方面。首先，监控 SIP 规范化。根据系统定义的 AIP 结构，利用 PREMIS 生成保存系统需要的元数据，包括描述、技术、保存元数据等；利用 JHOVE 检查 SIP 的格式和完整性，采用 PRONOM 进行格式注册，使用 Droid 自动批处理文件格式，生成规范化 SIP；检查规范化 SIP 的格式、结构、组成成分以及相关链接。其次，监控摄入过程中位流文件的完整性和功能性。系统为每个数据流生成 MD5 以进行一对多的一致性校验，同时验证文件内容可被指定程序读取并正确呈现。摄入完成后对新生成的 AIP 进行真实性和完整性验证并生成增量索引。最后，系统对新摄入且符合要求的 AIP 采用多备份、异质、异地备份机制进行备份，记录 AIP 永久标识符、存储路径、存储时间、提交人等信息。系统记录摄入行为相关信息，包括操作人、时间、摄入条目、成功/失败条目、备份位置等，发送摄入报告。

3.2 存档过程

3.2.1 AIP 系统应监控 AIP 的可读性、真实性和完整性。定期检查 AIP 在指定位置存在且可读；采用消息摘要的方法验证 AIP 和各种元数据的真实性；检测 AIP 声明包含的所有位流文件和元数据是

否存在，以保证 AIP 的结构完整性；检查 AIP 声明的位流数据和元数据之间的关系是否存在且正确完整，以确保 AIP 的关系完整性。系统还应监控 AIP 中的位流文件。根据存储信息检查文件存储位置的正确性、存储介质的可用性和数据的可读性；采用消息摘要验证位流文件的真实性；检查位流文件的完整性和功能性；监控位流文件格式，提取格式的相关信息并统计保存系统中各种格式的文件数量，以防止文件格式过时。

3.2.2 AIP 修改行为 在发生修改 AIP 的事件时，系统要详细记录事件的各个要素，包括标识符、类型、发生时间、细节、结果信息等属性。系统记录的事件属性元数据应作为原有 AIP 元数据的一部分被长期保存。系统在监控修改事件的同时，应保证原始文档的可追踪和修改过程的可逆性。可借鉴 Fedora 系统的版本管理特性，每次修改生成新版本，将数字对象的新版本与历史版本一起存储在保存系统中，完整记录数字对象的历史演变过程。

3.2.3 备份 每个 AIP 至少有 3 个备份数据且存储在不同地点的不同介质上。首先，确保备份 AIP 全部存在且可读。系统根据数据摄入成功后的备份信息找出 AIP 的全部备份数据并对其可读性进行验证。若有备份数据丢失或不可读，如缺失 AIP 永久标识符、时间、存储路径等信息，应立即补充并提交报告。其次，验证备份 AIP 的真实性和完整性。在确认源 AIP 通过真实性、完整性验证的基础上计算备份数据的消息摘要并将其与源 AIP 的消息摘要进行对比，如果相同则认为备份数据完好。如果不能确定源 AIP 是否正确，可参考 LOCKSS 的投票策略，分别计算每个备份 AIP 的消息摘要，得票多者为正确的 AIP。最后，保持 AIP 与其备份的一致性。在 AIP 修改前系统应验证 AIP 与其备份文件的一致性；保证其一致性后再对 AIP 及其备份进行修改，修改完成后再次进行一致性验证。

3.3 访问过程

3.3.1 访问行为 包括监控用户访问行为是否安全、访问功能是否正常，以及记录和分析访问日志。系统依据访问控制策略监控访问行为中的异常情况，如未授权用户操作数据或试图更改系统配置、用户进行恶意下载等，系统记录异常日志并发出告警。系统监控每次用户请求的响应结果和响应时间，如果出现大范围的获取失败或时间延迟，应立即通知管理员。系统记录用户访问行为的信息，如用户名、时间、行为、获取的信息、是否成功、IP 地址等，以便统计分析。

3.3.2 DIP 完整性和真实性 系统根据用户的请求清单对 DIP 的内容进行完整性检测，确保 DIP 包含用户请求的全部内容。同时系统对 DIP 中包含的每种元数据和文件与 AIP 中的元数据和文件做真实性校验，确保 DIP 全部元素的真实性。

3.3.3 数据机密性 如果访问过程涉及敏感数据，真实数据需要通过脱敏规则进行变形改造才能提供使用，对于重要数据采用加解密技术进行传输。系统需要监控脱敏处理和传输过程中数据的机密性。

3.4 系统运行环境

3.4.1 机房环境 通过在机房部署温湿度监控系统、空调监控系统、UPS 监控系统、漏水监测系统、消防报警系统和安全防盗系统，实现对机房温度、湿度、精密空调、UPS 电源、漏水、火灾、安全等情况的监测和报警。

3.4.2 硬件设备 包括监控服务器运行状态和 CPU、内存、磁盘等资源的使用情况，监控存储设备的硬盘状态、磁盘剩余空间、备份任务状态等信息，监控网络设备的温度、内存状态、端口可用性、端口流量、安全等实时情况；设置告警规则，提供性能阀值告警和故障告警。数据中心部署设备监控平台对硬件设备进行统一监控管理。监控平台对所有操作日志和系统运行日志进行监控，通过物理拓扑图和系统自动刷新直观展示硬件设备的实时运行状态，采用邮件和短信的方式向管理员发送告警信息和数据统计报告。

3.4.3 软件系统 系统需要监控操作系统、数据库、杀毒软件和 Web 应用程序等软件的使用期限、操作日志、启动和停止服务的时间和原因，定期对 Web 应用系统进行漏洞扫描。同时系统应监控长期

保存各个子系统的可用性、数据库状态和各种操作。

4 结语

医学数字资源长期保存系统需要构建本地化、可信赖的数字保存仓储，保存系统和保存对象的可信赖是其核心目标，监控管理能够为此提供很大保障。医学数字资源长期保存系统监控策略基于可信赖认证标准的要求制定，充分考虑到医学数字资源的特点和影响因素，覆盖数字保存生命周期的各个方面，可有效保证系统的稳定性和数字对象的不变性。目前数据中心已初步实现对长期保存系统运行环境的监控。保存系统建成上线后，将按照监控策略对整个系统的运行稳定性、安全性、处理流程、策略执行情况以及数字对象的真实性、完整性、可用性等方面进行监控，接收来自系统内部各个模块的报警信息和统计信息，将信息输出给相关系统和管理员。在未来工作中需要进一步研究长期保存系统的监控管理机制、技术和方法，探索基于人工智能的自动化监控。

参考文献

1 吴振新，付鸿鹄，马海收. 长期保存系统监控服务内容框架研究［J］. 图书情报工作，2014，58（3）：51－57，94.

2 The Consultative Committee for Space Data Systems. Reference Model for An Open Archival Information System［EB/OL］.［2017－10－13］. http：//public. ccsds. org/publications/archive/650x0m2. pdf.

3 The DCC. DCC Curation Lifecycle Model［EB/OL］.［2017－10－20］. http：//www. dcc. ac. uk/resources/curation－lifecycle－model.

4 University of Essex，University of Manchester and Jisc. Research Data Lifecycle［EB/OL］.［2017－10－20］. http：//www. data－archive. ac. uk/create－manage/life－cycle.

5 Data ONE. Data Life Cycle［EB/OL］.［2017－10－20］. https：//www. dataone. org/data－life－cycle.

6 CCSDC. Audit and Certification of Trustworthy Digital Repositories［S/0L］.［2017－10－15］. http：// public. ccsds. org/publications/archive/652x0m1. pdf.

7 Stanford University. LOCKSS［EB/OL］.［2017－10－18］. https：//www. lockss. org/.

8 Scott Prater. Fedora Digital Object Model［EB/OL］.［2017－10－18］. http：//wiki. duraspace. org/display/FEDORA38/Fedora＋Digital＋Object＋Model.

9 Ifaria. Scout——一个长期保存监测系统［EB/OL］.［2017－10－20］. http：//openpreservation. org/blog/2013/12/16/scout－preservation－watch－system/.

10 徐拥军，张倩. 加拿大图书档案馆的数字保存策略－可信数字仓储［J］. 档案学研究，2014（3）：90－96.

11 吴振新，王玉菊，付鸿鹄. 构建可信赖的数字资源长期保存系统摄入工作流［J］. 现代图书情报工作，2015，256（3）：1－7.

12 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.