

医学数字资源长期保存标准体系研究^{*}

杨晨柳 方 安 胡佳慧 王 茜

(中国医学科学院/北京协和医学院医学信息研究所 北京 100020)

〔摘要〕 针对医学数字资源构建长期保存标准体系,结合研究现状,设计体系框架,为具有长期保存价值的医学数字资源提供保存标准,以期指导后续保存系统应用实践。

〔关键词〕 医学数字资源;长期保存;标准体系

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2020.02.003

Study on the Long-term Preservation Standard System for Medical Digital Resources YANG Chenliu, FANG An, HU Jiahui, WANG Qian, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔Abstract〕 The long-term preservation standard system for medical digital resources has been established. The paper designs the system framework combined with current study situation, provides preservation standard for medical digital resources with long-term preservation value, in order to guide the application practice of subsequent preservation systems.

〔Keywords〕 medical digital resources; long-term preservation; standard system

1 引言

步入信息时代,数字资源已成为影响国家安全、长远发展以及人类文明传承的重要资源,其长期保存的重要性和紧迫性尤为突出。数字资源长期保存是指对目前社会所拥有和使用的数字形态的资源进行有效保存^[1],主要包含两层含义:一是长期

存储,二是长期可获取。随着大数据技术的发展,作为支撑数据密集型科学发现的重要资源,医学数字资源具有类型复杂、分布广泛、规模庞大、隐私要求高等特点^[2],存在切实的长期保存需求。

在数字资源生命周期的全过程中,标准是保证长期保存各环节互操作的基础。近年来医学领域的数字资源数量增长迅速且类型复杂多样,给存储标准化带来巨大挑战。针对医学数字资源特点,分析已有的长期保存基础框架,探索并梳理长期保存过程中涉及到的标准,完善和发展其标准体系框架,对医学领域资源存储标准化具有重要意义。为实现医学数字资源在长期保存过程中的有效组织、管理和检索服务,本文从研究现状、数字资源保存标准体系、医学数字资源保存标准建设3个方面研究并探索医学数字资源长期保存标准体系框架。基于层次化的分析与设计,重点围绕医学数字资源特点开展研究,探索出适用于医学数字资源长期保存标准体系建设方案,为系统建设提供支持。

〔修回日期〕 2019-07-08

〔作者简介〕 杨晨柳,助理研究员;通讯作者:胡佳慧,助理研究员。

〔基金项目〕 国家人口与健康科学数据共享服务平台“长期保存系统建设”后补助项目(项目编号:NCMI-KD07N-201905);中国医学科学院医学与健康科技创新工程服务“一带一路”战略先导科研专项“卫生信息服务研究”(项目编号:2017-I2M-B&R-10)。

2 研究现状

开放档案信息系统 (Open Archival Information System, OAIS)、保存元数据实施策略 (Preservation Metadata Implementation Strategies, PREMIS)、元数据编码和传输标准 (Metadata Encoding and Transmission Standard, METS) 等国际或行业通用标准为长期保存存档概念提供指导性框架。其中 OAIS 是由美国空间数据系统咨询委员会 (Consultative Committee for Space Data System, CCSDS) 制定的标准, 于 2003 年被作为 ISO 标准 (ISO 14721) 发布^[3], 为长期保存系统建立基本概念和参考模型且在数据资源长期保存领域具有广泛的影响力; PREMIS 是联机计算机图书馆中心 (Online Computer Library Center, OCLC) 和研究图书馆组织 (Research Library Group, RLG) 在 2003 年发起的保存元数据实施战略^[4], 主要用于描述有关数字资源保存特征信息; METS 由数字图书馆联盟发起, 为数字图书馆保存系统提供数字对象和交换元数据的编码格式, 用于元数据编码、建立关联和数据迁移。数字资源长期保存标准体系建设方面, 郭家义^[5]将标准分为系统、业务逻辑和数据层次; 刘振^[6]将体系框架从流程和层次结构两个角度进行规划和管理; 王乐春等^[7]结合国家图书馆实际情况提出遵循 OAIS 框架并兼容 PREMIS 和 METS 的长期保存数据组织标准。此外国内部分机构针对自身文献资源存储类型也制定相应的数字资源长期保存元数据标准, 如北京大学图书馆的《中文元数据标准框架》^[8-9]、清华大学图书馆的保存元数据方案^[10]等。医学数字资源种类繁多且具有长期保存的切实需求^[2]。OAIS 等作为长期保存系统的参考标准, 在面向具体的数据保存应用场景中应结合实际需求加以利用。为实现医学文献资源的长期保存, 解放军医学图书馆秦将华^[11]提出基于 OAIS 的保存元数据框架。医学数字资源长期保存系统建设过程中为实现对医学大数据的长期保存, 需满足系统建设、业务发展和数据保存等各方面需求。因此本文针对医学数字资源构建长期保存标准体系, 为具有长期保存价值的医学数

字资源提供保存标准的建设思路, 以期指导后续保存系统的应用实践。

3 医学数字资源保存标准体系

3.1 概述

围绕 OAIS 框架, 基于系统实现角度将医学数字资源长期保存标准体系框架分为系统标准、业务标准和数据标准 3 部分, 见图 1。

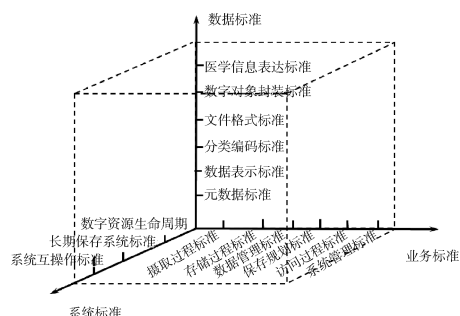


图1 医学数字资源保存标准体系

3.2 系统标准

3.2.1 系统参考模型 构建医学数字资源长期保存系统的关键在于遵循统一的数字信息保存标准, 从而确保系统中资源和服务的可操作性与共享性。OAIS 为长期保存系统创建的数字资源存取基本概念框架, 现已成为国际上数字资源长期保存研究与开发项目有限遵循的标准规范。

3.2.2 系统互操作标准 卫生信息领域有常用的系统互操作标准, 如医疗数据交换标准 HL7^[12] 和医疗健康信息集成规范 (Integrating the Health Enterprise, IHE)^[13]。HL7 支持使用现行的各种医学编码标准, 方便各个医疗机构的应用系统间能够进行医学资源的互通; IHE 通过采用卫生信息标准, 促进卫生信息在系统间、机构间实现无缝传递, 同时基于已有的医疗信息领域规范和标准 (如 DICOM、HL7、ebXML 等) 对不同层次医疗服务进行优化, 有助于推动各医疗信息系统间实现资源共享。

3.2.3 数字资源生命周期模型 生命周期模型是生命周期管理的基础, 基于不同形式的资源 (文本、音频、视频、图像等), 采用合适的模型与方

案,对被保存的医学数字资源的生命周期管理有重要意义。英国数字资源管理中心(Digital Curation Center, DCC)^[14]于 2008 年发布数字资源保存生命周期模型,描述数字对象产生、管理、保存的各个环节,强调通过对整个保存生命周期的管理来维护数字对象的可信赖性、可用性、完整性。

3.3 业务标准

3.3.1 数据摄取 主要指从出版商或其他信息提供者收集或接收统一组织格式的提交信息包(Submission Information Package, SIP)并对其进行检查、转换、规范和处理等。摄取标准用于规范被接收到的信息内容。目前摄取领域的标准主要是生产档案标准接口方法(Producer Archive Interface Methodology Abstract Standard, PAIMAS)^[15],旨在识别、定义和提供生产者和存档之间的关系和交互结构。此外摄取过程还涉及安全监测、完整性校验等方面的标准。安全监测主要对 SIP 进行病毒、木马、恶意代码扫描,消除安全隐患;完整性校验普遍采用信息加密技术,常用的算法标准有 MD5、SHA 等。

3.3.2 数据存储 数据存储提供对存档信息包(Archival Information Package, AIP)的存储、维护和检索功能与服务,子功能包括从摄取实体接收 AIP 并将其添加到存储器、管理存储体系、更新存储介质、例行错误检查、灾难备份、为检索模块提供 AIP 等。存储过程中的标准规范主要指导保存的过程、方法、技术等方面的问题,主要包括内容对象标准、数据编码标注、数字迁移标准、存储介质和方式标准等^[5]。

3.3.3 数据管理 主要提供填充、维护和存取用于标识记录库存资源的描述性信息与管理文档的管理型数据。数据管理标准用于管理数据库(维护数据库模式、视图及系统完整性等)、执行数据库更新(加载的描述信息或档案管理数据)、提供数据查询、产生结果等。

3.3.4 保存规划 侧重于 OAIS 环境的监测服务,确保被存储的信息在很长一段时间后,即使原有保存环境发生变化资源仍然能够被相应的目标用户团体所存取。由于医学数字对象的长期保存主要依赖

各类物理存储设备,如果不能监测并及时处理设备出现的问题,将直接导致资源的损坏乃至丢失。

3.3.5 数据访问 主要与用户联系以接收用户请求,运用访问控制标准来限制对某些特别保护信息的访问,严格把控请求的执行以成功完成响应并传递信息给用户。被保存信息的访问过程主要涉及检索、访问控制与授权等方面的标准^[6]。

3.3.6 系统管理 在医学数字资源长期保存过程中存在各类安全隐患,包含物理环境、基础设施等,直接威胁被保存数字资源的安全。信息安全管理标准(ISO 27001)建立科学有效的管理制度来保障数字资源的机密性、完整性和可靠性。同时可信赖仓储认证对构建医学数字资源存储体系也具有重要意义,制定可信任数字仓储的认证标准是建立可信任仓储系统的核心内容。

3.4 数据标准

3.4.1 医学信息表达 目前国际卫生信息领域常用的医学术语标准主要有《国际疾病分类》(International Classification of Diseases, ICD)^[16]、临床医学系统术语(Systematized Nomenclature of Medicine - Clinical Terms, SNOMED CT)^[17]、Read 临床分类(Read Clinical Classification, RCC)^[18]、统一医学语言系统(Unified Medical Language System, UMLS)^[19]等。其中 ICD 涉及所有疾病和死亡原因,具有权威性、科学性并宏观反映居民健康情况特征;SNOMED CT 形成完整的医学术语体系,涵盖大多数方面的临床信息;RCC 覆盖病历中的所有术语;UMLS 对生物医学科学领域内已有的多个受控词表进行汇编,提供词表之间的映射,因此 UMLS 也被看作是生物医学概念所构成的本体和叙词表。

3.4.2 数字对象封装 在医学数字资源长期保存中,信息包的选择、封装和存档是关键问题,数字对象封装标准用于规范信息包的组织和构成。长期保存项目通常采用 METS^[20]、JATS^[21]和 RDF^[22]等方式来实现对信息包的封装。其中 METS 在为数字资源长期保存提供封装模型的同时也为系统的构建提供数据模型;JATS 标签集提供一组定义元素和属性的 XML Schema 模型来描述期刊论著及非论著材

料并已经在出版商、科技期刊以及图书馆等领域广泛使用; RDF 是 W3C 推荐使用的描述资源及其之间关系的规范, 能够编码、交换和重用结构化元数据, 其通过支持语义语法和结构的设计提升元数据的互操作性^[23]。

3.4.3 文件格式 有效创建、交换和保存数字对象信息的基础, 在医学数字资源长期保存过程中采用开放标准的格式有利于对资源进行创建和维护, 也有利于实现系统之间的互操作。长期保存研究实践中各类机构或组织根据自身需求和特点, 形成较为成熟的格式使用规范, 如美国佛罗里达数字档案项目、美国国会图书馆、英国数据档案的推荐格式, 这些规范从各角度归纳和总结 PDF/A、JPEG、Tiff、MPEG 等重点格式的发展、特点和应用, 经过实践考验, 值得研究借鉴。

3.4.4 分类编码 数据编码主要用于对被保存数字对象的描述与识别。在卫生信息领域已有部分临床信息系统可采用、能标识各种实验室检测项目和临床观测指标的编码标准。世界卫生组织 (World Health Organization, WHO) 制定国际统一的疾病分类方法 ICD; 多数西方国家的政府卫生部门和商业性医疗机构认可并采用的观测指标标记符逻辑命名与编码系统 (Logical Observation Identifiers Names and lode, LOINC)^[24]; 对 ICD 有更全面的细分, 不仅含有诊断编码, 还含有就诊原因、治疗原因和实验结果代码的国际社区医疗分类 (International Classification of Primary Care, ICPC)^[25]。就我国而言, 卫生信息数据集分类与编码规则 (WS/T 306 - 2009)^[26]规定卫生信息数据集分类与编码需遵循的基本原则、技术方法以及应用规则, 适用于医药卫生领域各类卫生信息数据集的分类与编码方案。

3.4.5 数据标识 对长期保存的各类医学数字对象 (如文献、数据集、医学影像等) 进行唯一标识, 可以方便快速地定位和访问这些数字资源, 将资源与其他作者实体、机构、项目或研究团队相关联, 便于发现和引用同被保存的其他医学数字资源。传统出版物如书刊、磁带、光盘都有国际标准编号 (ISBN、ISSN、ISCN) 及其条形码作为出版物的唯一标识; 常见的数字对象唯一标识符有 DOI、

SICI、BICI 和 PII 等^[5]。在卫生信息领域中, LOINC 为实验室和临床检查提供一套统一的名称和标识码, 每个概念具有唯一代码且恒久不变。目前 LOINC 在国内外得到广泛认可并在许多卫生医疗机构得到广泛应用。

3.4.6 元数据 实现和支撑数字资源长期保存过程及相关操作的基础, 也是保证医学数字资源有效性、可操作性和可理解性的重要信息。元数据对数字对象进行描述, 同时在保存过程中对数字资源进行识别并管理, 实现数字资源的发现、查找、共享和管理。根据描述对象的差别元数据可被分为技术、描述、保存和管理元数据等。在医学数字资源长期保存过程中元数据方案的开发应用涉及的问题主要包括元数据的应用领域和目标、内容结构、编码语言、互操作和检索体系以及长期保存等。

4 医学数字资源保存标准建设

4.1 系统标准

4.1.1 基于 OAIS 的系统架构 医学数字资源长期保存系统框架沿用 OAIS 参考模型。该模型作为致力于维护信息系统中数字信息的长期保护和存取的参考模型和基本概念框架, 现已成为国际上数字资源长期保存研究与开发项目有限遵循的标准规范, 被应用于 DSEP、Cedars、DIAS、CDS、DAITSS 等具有代表性的数字资源长期保存项目中^[27]。因此在 OAIS 基础框架下开发医学数字资源长期保存系统能够更好地实现医学领域间的资源共享和系统间的交互。

4.1.2 基于保存生命周期的业务流标准 基于数字资源生命周期全过程设计标准体系框架。由于医学数字资源长期保存涉及数字资源生命周期的全过程, 其中各流程都需要标准的支持, 因此系统地研究医学数字资源长期保存过程中的相关标准对于系统的设计和实现是十分必要的。

4.1.3 基于医疗卫生通用互操作标准 系统互操作协议在通用标准和长期保存业界专有标准的基础上采用 IHE 作为系统互操作设计基础, 同时借鉴 HL7 进行交互方案的设计。其中 HL7 在医院信息系统 (Hospital Information System, HIS) 接口基础上

定义了接口标准格式且支持使用现行的各种编码标准,如 ICD-9/10、SNOMED CT 等。IHE 可有效推动不同医疗信息系统间实现信息共享。

4.2 业务标准

4.2.1 基于 OAIS 的功能模型 各子模块创建业务标准体系。在系统建设方面,OAIS 功能模型将数字资源长期保存系统划分为摄取、存储、数据管理、保存规划、访问、系统管理 6 个部分,各部分都需要对应的指导与实践标准。采用标准化的协议和规范,可以为医学数字资源长期保存系统建设提供科学依据和指导,从而确保系统各功能模块的实现。

4.2.2 细化业务标准 业务标准包含数据和系统管理标准。由于数据和系统管理都与管理相关,因此在已有的一些数字资源长期保存标准体系中将数据管理和系统管理合并为管理标准。事实上在 OAIS 功能模型中,数据管理与系统管理被分为两个独立模块。其中数据管理模块包含便于发现和存取被保存资源的描述元数据和管理元数据;系统管理模块用于系统的日常检查、运维和访问等操作。比较发现两者功能及侧重点并不完全一致,因此将二者分别列出并加以说明。

4.2.3 增加保存规划标准 保存规划提供的服务和功能是检测 OAIS 的环境,保证资源的长期可存取。由于医学数字资源的长期保存主要依赖各类物理存储设备,如果不能监测并及时处理设备出现的问题将直接导致被存储资源的损坏乃至丢失,因此将保存规划标准列入业务标准中,用于维护系统整体保存环境。保存规划标准主要包含存档硬件安全管理和机房环境监控两部分内容。

4.3 数据标准

4.3.1 增加医学信息表达标准 相较于文献数字资源,医学数字资源具有层次性和专业性强、隐私性要求高等特点,因此在开展医学科学数据的长期保存活动时需在遵从医学信息表达相关标准规范的前提下进行。目前国际卫生信息领域常用的医学信息表达标准主要有 ICD、SNOMED CT、UMLS、RCC 等,可依据具体存储及检索需求对上述医学信

息表达标准进行选择及复用。

4.3.2 选用 RDF 描述框架 作为被保存医学数字资源的封装标准。在医学数字资源长期保存过程中对数字对象进行封装可以有效解决信息包内各类信息的构成和组织问题,使数字资源被更好地保存。医学数字资源长期保存系统采用 Fedora 4.0 版本作为底层仓储以 RDF 存储数字资源,可以形成语义数据,方便进行语义挖掘、推理,构建关联数据平台。RDF 描述框架逐步在国际上得到广泛认可与应用。

4.3.3 增加医学影像类型 在文件格式标准中添加医学影像类型及推荐格式 DICOM。在长期保存实践中各类机构或组织根据自身需求和特点,各自形成一套较为成熟的格式使用规范。相比较而言,医学数字资源类型丰富、格式多样,除文本、图表、压缩文件外还包括事实数据集(库)和实验数据集(库)以及医学检查所产生的医学影像文件。DICOM 的应用范围包括 CT、MRI、DSA、SPECT、超声、数字化 X 线影像(CR、DR、FPD)以及内窥镜、病理学、耳科、皮肤和中医舌苔图像等,因此在文件格式标准中加入 DICOM 可以更全面地实现被保存资源的转换和存储。

4.3.4 增加医学数据编码标准 由于医学信息资源的专业性,在原有文献数字资源编码基础上添加医学数据编码标准。在卫生信息领域中已有部分可以作为所有临床信息系统可采用、能够标识各种实验室检测项目和临床观测指标的编码标准,如前文提到的 ICD、LOINC、ICPC 及 WS/T 306-2009 卫生信息数据集分类与编码标准。

4.3.5 构建元数据框架 构建包含技术、保存、描述、管理元数据 4 部分内容的元数据。基于 OAIS 参考模型的信息模型可以抽象出一个 3 层架构的元数据框架,第 1 层是关于内容数据呈现的元数据,即表现信息部分,对应描述元数据;第 2 层是保存描述信息部分,可记录长期保存过程中医学数字资源所处的环境信息,对应技术元数据和保存元数据;第 3 层是系统管理层面的元数据内容,实现对数字资源加工过程的管理等,对应管理元数据。

5 结语

通过对数字资源长期保存标准体系和科学数据平台元数据标准研究现状的分析,结合医学数字资源自身特点和长期保存需求,本文提出医学数字资源长期保存标准体系。基于数字资源长期保存已有的标准体系,将医学数字资源特有内容进行扩展和延伸,设计涵盖系统、业务、数据标准的体系框架。在该宏观标准体系的指导下,结合 Fedora 4.0 底层仓储结构,目前医学数字资源长期保存系统 MedPRES 已完成由摄取到管理的业务流程功能建设。同时针对系统建设中的实际问题及时调整对应标准和建设方案,以指导 MedPRES 的后续完善,从而适应实际保存需求,为实现医学数字资源长期保存系统中保存对象的长期可用奠定基础。

参考文献

- 谷紫藤. 我国数字资源长期保存策略研究 [J]. 现代营销: 学苑版, 2017 (7): 290-290.
- 胡佳慧, 方安, 杨晨柳. 医学数字资源长期保存存储策略研究 [J]. 医学信息学杂志, 2017, 38 (12): 73-78.
- 张鲁冀. OAIS 参考模型功能及其应用研究 [J]. 天津科技, 2015, 42 (7): 59-60.
- 高嵩, 张智雄. PREMIS 保存元数据体系分析 [J]. 现代图书情报技术, 2006 (4): 19-23, 52.
- 郭家义. 数字信息资源长期保存系统的标准体系研究 [J]. 数据分析与知识发现, 2016, 1 (4): 14-18.
- 刘振. 数字资源长期保存的标准体系研究 [J]. 情报理论与实践, 2016, 39 (7): 45-49.
- 王乐春, 宋姝, 董晓莉. 超大型数据资源长期保存系统数据组织标准研究 [J]. 标准科学, 2012 (8): 12-15.
- 王松林. 《中文元数据方案》的特点及有关问题的思考 [J]. 数字图书馆论坛, 2005 (12): 22-27.
- 肖珑, 陈凌, 冯项云, 等. 中文元数据标准框架及其应用 [J]. 大学图书馆学报, 2001, 19 (5): 29-35.
- 牛金芳, 郑小惠, 曾婷, 等. 清华大学图书馆保存元数据方案 [J]. 大学图书馆学报, 2003, 21 (2): 22-25.
- 秦将华. 医学图书馆数字资源长期保存技术应用框架研究 [D]. 北京: 中国人民解放军军事医学科学院, 2011.
- HL7 International. Health Level Seven [EB/OL]. [2019-04-08]. <http://www.hl7.org/>.
- IHE International. IHE [EB/OL]. [2019-04-08]. <https://www.ihe.net/>.
- WIKIPEDIA. Digital Curation Centre [EB/OL]. [2019-04-08]. https://en.wikipedia.org/wiki/Digital_Curation_Centre.
- Consultative Committee for Space Data Systems. Producer Archive Interface Methodology Abstract Standard (PAIMAS) [EB/OL]. [2019-04-08]. <https://public.ccsds.org/Pubs/651x0m1.pdf>.
- CDC. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) [EB/OL]. [2019-04-08]. <https://www.cdc.gov/nchs/icd/icd10cm.htm>.
- SNOMED International. SNOMED CT [EB/OL]. [2019-04-08]. <https://www.snomed.org/snomed-ct>.
- NHS. Read Coded Clinical Terms [EB/OL]. [2019-04-08]. https://www.datadictionary.nhs.uk/web_site_content/supporting_information/clinical_coding/read_coded_clinical_terms.asp?shownav=1.
- U. S. National Library of Medicine. Unified Medical Language System (UMLS) [EB/OL]. [2019-04-08]. <https://www.nlm.nih.gov/research/umls/>.
- The Library of Congress. METS [EB/OL]. [2019-04-08]. <http://www.loc.gov/standards/mets/>.
- National Center for Biotechnology Information. JATS [EB/OL]. [2019-04-08]. <https://jats.nlm.nih.gov/>.
- W3C. RDF [EB/OL]. [2019-04-08]. <https://www.w3.org/RDF/>.
- 柳亚琴, 温彩玲, 余雪丽, 等. RDF 在语义 Web 知识表示中的作用 [J]. 太原理工大学学报, 2006, 37 (2): 146-149.
- LOINC Committee. LOINC [EB/OL]. [2019-04-08]. <https://loinc.org/>.
- U. S. National Library of Medicine. ICPC [EB/OL]. [2019-04-08]. <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ICPC/>.
- 中国公共卫生科学数据中心. 卫生信息数据集分类与编码规则 [EB/OL]. [2019-04-08]. <http://www.doc88.com/p-6671631996924.html>.
- 张智雄, 郭家义, 吴振新, 等. 基于 OAIS 的主要数字保存系统研究 [J]. 现代图书情报技术, 2005 (11): 1-9, 13.