基于OCFL的数字资源保存文件系统设计*

姚宽达 方安 杨晨柳 王蕾 胡佳慧
（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：本文针对数据密集型科研环境下的科研数据长期保存需求，以牛津通用文件布局方法（OCFL）为基础，设计支持对象存储及版本控制的数字资源保存文件系统，并在医学大数据长期保存系统环境中进行了应用实现和效果分析。

关键词：数字保存；OCFL；数据密集型科研；文件存储；版本控制
中图分类号：G250　DOI：10.3772/j.issn.1673-2286.2021.12.009
引文格式：姚宽达，方安，杨晨柳，等. 基于OCFL的数字资源保存文件系统设计[J]. 数字图书馆论坛，2021（12）：58-64.

信息技术的飞速发展使得数据获取难度日益降低，大数据时代，科学研究已向数据密集型计算科学[1]转变，数据对于科学研究的重要性显著提高。数据密集型科研环境下，科学研究和创新活动越来越依赖于对大规模数据的分析、挖掘以及再利用[2]，由此产生的科研数据和成果具有较高的保存价值，是长期科研过程中的重点保存内容[3]。数字资源长期保存是对数字内容进行持续管理和维护的一系列活动，要求在长期保存数字资源的同时，确保保存信息的真实可信，且能够被未来的使用者所理解和利用[4]。相较于一般数据，科研数据不仅体量大、更迭快、类型杂，还具有学科差异性、知识关联性以及历史积累性等典型特征[5]，这给长期保存的数据完整性及内容连续性等方面带来一系列挑战。因此，针对数据密集型科研环境下的科研数据特征，本文在分析科研数据长期保存需求的基础上，设计数字资源保存文件系统，并结合实际应用开展相关探索。

1 研究现状及意义

1.1 现有的数字保存存储策略分析

技术环境的不断革新给数字保存带来诸多挑战，为避免由此导致的数字内容不可用等问题，资源保障机构需要不断更新数字仓储的保存策略以适应新形势下的数字保存需求[6]。开放档案信息系统（Open Archival Information System，OAIS）[7]模型为数字保存提供了指导性原则，对保存内容文件和保存描述信息进行了界定，并提出了信息包的概念，但该模型没有为保存系统的构建和实际应用给出具体的保存方法和建议[8]。因此，面向特定的数字保存需求，保存机构与学者开展了相关研究和探索，提出和制定了一系列存储策略和保存方案，其中文件存储和版本控制是数字保存系统设计的关键要素。美国国家进化综合中心与北卡罗来纳大学等5家机构合作开发了Dryad数据库[9]，其存储策略支持版本控制，但不建议对已提交数据的频繁更新[10]。斯坦福大学图书馆使用Moab方法同样支持版本控制，但其设计缺乏广泛的适用性[11]。诺特丹大学赫斯堡图书馆在保存系统中使用的BagIt方法策略在文件传输上效率较高，但不支持版本控制[12]。北京大学图书馆探索了数字长期保存系统（Digital Preservation System，DPS）[13]在高校图书馆中的应用与服务，但存储系统对于底层数据模型兼容型的支持不够充分[14]。

* 本研究得到国家社会科学基金项目“突发公共卫生事件网络信息资源的知识图谱构建研究”（编号：21CTQ016）资助。

1.2 数据密集型科研环境下数字保存的特征与意义

数据密集型科研环境下的科研数据一方面具有大数据的一般特征，如数据规模庞大、来源分散、数据结构多样以及具有研究和使用价值等；另一方面，相较一般的大数据，科研数据还具有以下典型特征。

（1）学科差异性。不同学科领域的科研数据在数据结构、文件类型甚至数据体量等方面具有较大的差异。

（2）知识关联性。科研数据更强调从大量、积累的科研数据中演化和发现新的科研规律或知识[5]，对时效性要求较低。

（3）历史积累性。科研数据更注重数据的历史积累和数据保存体系的完整性，需要对历史积累的数据进行重新分析。

（4）数据价值性。科研数据需要在保证数据准确和完整的前提下，其研究和利用价值才能得到体现[15]。

相较于传统科研范式，数据密集型科研环境下的数据分析和知识发现不再依赖严密的假设检验过程[16]，而是通过对跨时间、跨空间、跨领域的更大规模的科学数据循环进行采集、分析及存储，使数据成为科研的对象和工具，以此为基础产生新的科学研究方法。数据密集型科研环境下，数据需要不断地被重用和验证，因此，作为科研数据生命周期的重要环节，数字保存对于有效保障和促进科学研究持续开展具有积极意义。

2 OCFL及其优势分析

针对数字仓储对数字对象通用储存方法的迫切需求，康奈尔大学、斯坦福大学、DuraSpace、牛津大学以及埃默里大学等机构于2018年共同提出牛津通用文件布局方法（Oxford Common File Layout，OCFL）[11, 17]。OCFL以结构化、透明化和可预测的方式对数字对象进行规范化存储，降低数字对象在存储结构上对应用程序的依赖；并使用正向增量的版本控制方法对数字对象的版本进行管理和溯源，以提升OCFL解决数字资源存储的迭代、冗余以及存取效率问题的能力。具体而言，OCFL的关键目标[11]表现在：①完整性，具备在没有额外信息资源的情况下重建存储库的能力；②人机可解析性，以确保在没有原始软件的情况下可以理解内容；③鲁棒性，针对错误、损坏和技术迁移的稳健存储性能；④版本控制，持续记录保存对象的历史信息以支持保存对象的更新和更改；⑤可扩展性，具备将内容存储在各种存储基础架构上的能力。

针对数字对象频繁更新带来的文件存储管理问题，OCFL数字对象将内容文件按照版本保存，从而减少对保存内容的读取操作，降低资源存储和重建成本，提高读取效率，增强管理便捷性。OCFL通过规范化的文件存储结构提升数字资源的可读性，确保管理者和应用程序能够快速识别文件布局，实现存储资源查询、检索、解析等一系列操作。针对系统软件和架构的变化以及数字资源内容的迁移，OCFL可以在数字仓储功能不完整的情况下，基于保存管理文件内容理解存储结构，结合简单的应用程序进行管理[18]。相较于直接使用长期保存应用系统进行存储，应用系统的更新对使用OCFL作为数字存储方法的长期保存系统的影响更小，降低了跨度数十年的长期保存活动中系统变化导致的底层存储修改的成本。

此外，OCFL使用基于正向增量的版本控制方法跟踪管理数字对象的历史，更加高效地重建历史版本，为数字对象版本信息的溯源提供保证。相较于全量版本对每个版本的内容文件分别进行全量存储，增量版本控制仅储存版本之间发生更改的文件，可以有效地减少存储的数据重复，降低存储空间的压力。正向增量版本控制法在添加新版本时更为简便，但在重建最新版本方面需要更多的资源和工作[19-20]。OCFL使用内容寻址技术结合保存管理文件对正向增量版本控制方法进行了改进，不仅利用内容寻址技术中的文件校验和值进行重复文件存储的消除[21]，还将其作为文件在存储系统中的标识符和定位器，从而使得OCFL在解决存储冗余的同时，降低了存储及重建时的消耗。

综上，OCFL满足数字资源存储对完整性、人机可解析性、鲁棒性和版本控制的需求，其以数字对象为基础的版本控制方法可以支撑科研数据的长期存储，规范化存储结构增加可读性，为迁移和重建等文件读写操作提供支持；正向版本控制方法在节省存储空间的同时还可支持数据的溯源，面对历史积累下形成的多版本科研数据拥有较好的适配性。除此之外，版本控制和人机可解析性防止了人为和非人为双方面的错误所导致的文件损坏，增强了整个文件系统的鲁棒性，保障了数据资源的准确性。

3 基于OCFL的数字资源保存文件系统设计

3.1 MedPRES保存需求分析

面向大数据环境下的医学数字资源长期保存需求，中国医学科学院医学信息研究所已开展医学大数据长期保存系统（MedPRES）建设[22]。MedPRES以信息包形式保存资源，并基于全量版本控制方法对保存资源进行统一管理。面对保存数据的持续更新需求，尤其是内容大量重复的新增数据，全量版本控制方法将导致存储冗余问题，其对多版本数据的存取与溯源较为复杂，增加系统管理成本。为提高MedPRES对科研数据的保存能力，数据密集型科研环境下的数字保存具有以下需求。

（1）以数字对象为单位进行管理。针对不同类型和具有不同文件结构的科研数据，采用以对象为单位的保存方式可有效应对科研数据间的差异性，为数字对象的通用性管理和版本控制提供便利。

（2）具备较强的版本控制能力。科研数据具有迭代更新的需求，为支持数据的跨时间多次存入，数字保存不仅需要保证保存内容的不变性，还要确保数字对象可溯源，较强的版本控制是科研数据连续性和完整性的重要保证。

（3）拥有良好的文件存储结构。在长期的科学研究过程中，数据需要多次存入和读取，良好的文件存储结构可有效降低科研数据的保存、取用和管理开销。

鉴于OCFL在对象存储和版本控制方面的优势，本文基于OCFL的存储文件结构进行数字资源保存文件系统的设计，以优化MedPRES的版本控制能力及文件存取效率，降低数字保存的存储空间和应用管理成本。

3.2 保存文件系统结构设计

基于上述对数据密集型科研环境下数字保存的需求分析，设计数字资源保存文件系统，如图1所示。

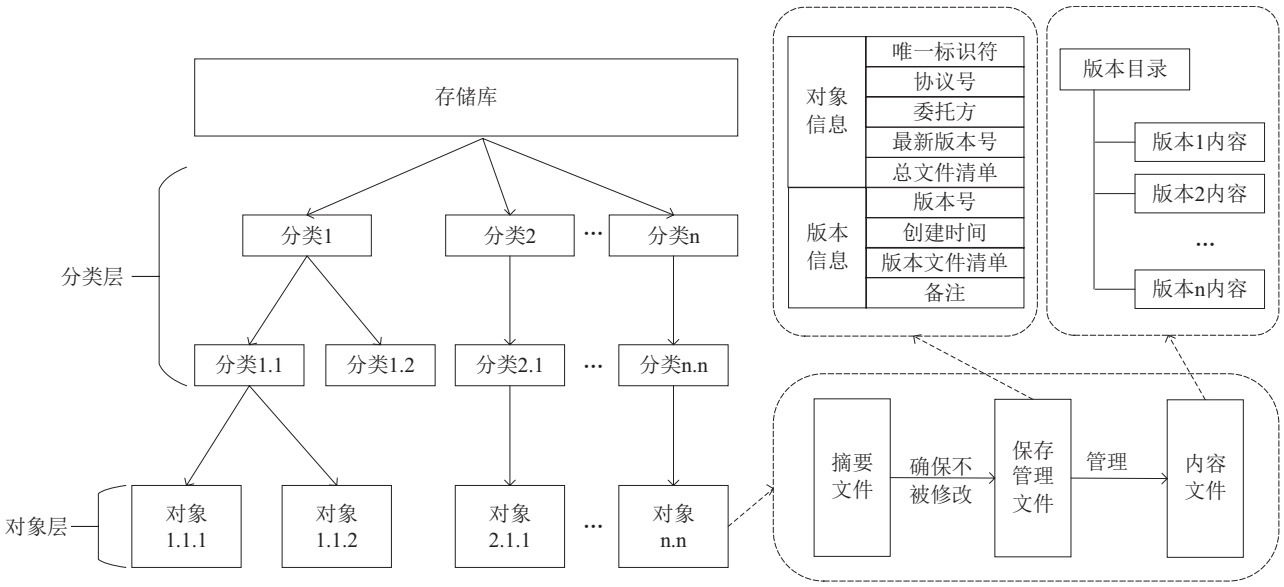

图1 数字资源保存文件系统框架

数字资源保存文件系统由分类层和对象层组成。分类层用于保存对象的分类存储和管理，第一级对应数据委托方信息，第二级对应数据类型（如科技文献或科学数据等）；对象层由数字资源保存对象组成，每个数字对象只属于一个分类。数字资源保存对象是一个由文件组成的树状层次结构，也是构成保存文件系统的重要部分。每个树状层次结构最上级的位置为对象根目录，对象根目录在数字仓储中由分类层节点和对象层节点链接形成的路径为该对象在保存文件系统中的URI，每个数字对象都对应一份数据保存协议，同一份协议下多次保存的数据都属于同一个数字对象包含范围内的数字资源。

数字对象中包含3个部分，即摘要文件、保存管理文件及内容文件。其中，内容文件以版本目录的形式存

储，保存管理文件用于管理内容文件，摘要文件确保保存管理文件在每次版本变化之间不会被修改。保存管理文件中的对象信息用于标识和管理数字资源保存对象的基本信息，是对保存对象的描述；版本信息用于管理各个版本的保存内容，每个版本存储的内容文件对应一份版本信息进行管理，是对版本的描述记录。

图2展示了包含一个版本信息的对象结构。其中，preservation是保存管理文件，用于管理和记录版本目录中各文件的结构内容和版本变化。preservation.digest是不变性校验文件，其中存储了preservation文件对应的校验值，用于保证保存管理文件的数据真实性和不变性。版本文件夹v1属于版本目录，其中包含初始版本的内容文件及保存管理文件，为版本控制提供文件结构层级的支持。

图2　数字对象文件结构示意图

数字对象根目录中的preservation文件为数字对象当前版本的保存管理文件，其中记录了当前版本中所包含的保存文件信息及其对应的管理信息。对象根目录的preservation文件与该对象最新版本子目录下的preservation文件完全相同。在对象根目录重复存储的保存管理文件位置固定，可减少查询最新版本对于版本目录的遍历操作开销。将对应版本的新增资源文件储存在该版本的版本文件夹下，可在添加新版本时不改变原有文件结构，只需添加新的版本文件夹并修改保存对象根目录中的保存管理文件，以尽可能地避免对数字资源的保存位置进行操作。对于preservation.digest文件中计算校验值的摘要算法可以根据系统需求进行选择，校验值的更新为数字对象版本变化的最后一步，以确保上一次版本变化结束到下一次版本变化开始前保存管理文件的不变性。

3.3　内容文件存储策略

保存文件系统采用基于OCFL的正向增量版本控制方法对数字对象进行管理。保存管理文件包括总文件清单和版本文件清单，其中，文件清单由保存文件的校验和及对应文件的保存路径或相对路径组成，校验和用于标识数字对象中的文件。总文件清单记录数字对象所有版本中包含的文件信息和实际保存路径，版本文件清单记录相应版本包含的文件信息和文件结构。通过文件清单的引入，保存对象的版本恢复不再需要依次遍历各个版本，而是可以根据版本清单中的记录直接进行版本恢复，从而避免版本重建成本过高问题。保存管理文件中应包含的基本项如表1所示。

表1　保存管理文件内容

信息块	基本项	描　述
对象信息	唯一标识符	该保存管理文件对应的保存对象唯一标识符
	协议号	该数字对象对应MedPRES系统中的协议ID
	委托方	声明该数字对象权利所属的委托方
	最新版本号	声明该数字对象的最新版本号
	总文件清单	由文件保存路径与文件校验和成对组成，包含该数字对象中所有版本的全部文件
版本信息	版本号	包含版本号、创建时间、文件列表和备注这4项
	创建时间	新版本文件入库的时间
	版本文件清单	由文件相对路径与文件校验和成对组成，仅包含该版本下的所有文件，文件校验和的值应包含在总文件清单的列表中
	备注	用于填写版本更新的相关信息，如修改内容或提交者信息等

对象信息部分的唯一标识符和最新版本号都用于描述当前保存对象的基本信息状态。通过唯一标识符确定对应的保存对象，一般为对象根目录的路径；通过最新版本号确认保存对象版本状态，从而节省遍历版本目录获取数字对象版本状态的成本。协议号与委托方信息用于标识数字对象的权利所属，保存协议中记录有保存内容的许可范围及保存时限等信息。

图3给出了数字对象结构的示例，该数字对象的初

始版本v1包含3个文件，版本v2将文件ch3.pdf重命名为ch4.pdf，并添加与原ch3.pdf完全不同内容的同名新文件ch3.pdf。版本v3对ch2.pdf文件进行了删除操作。版本v2中新增的文件ch3.pdf与数字对象原有的文件校验和值不同，而版本v1中的ch3.pdf只进行重命名操作但未改变文件内容，所以按照存储策略，版本v2的子文件夹中只保存新增的ch3.pdf文件。版本v3因不涉及新增文件，无须在版本文件夹中保存新的文件，只需要新增并更新保存管理文件。

```
[对象根目录]
    ├── preservation
    ├── preservation.digest
    ├── v1
    │    ├── preservation
    │    └── containfile
    │          ├──ch1.pdf
    │          ├──ch2.pdf
    │          └──ch3.pdf
    ├── v2
    │    ├── preservation
    │    └── containfile
    │          └──ch3.pdf
    └── v3
         └── preservation
```

图3 数字对象结构示意图

总文件清单仅以文件校验和作为文件标识符，不同版本的ch3.pdf文件对于文件系统而言是拥有不同校验和的不同版本子目录下的ch3.pdf同名文件。同一文件在总文件清单和版本文件清单中存储的路径各不相同。总文件清单中存入的是保存路径，保存路径是指文件相对于该保存对象根目录的文件路径。版本文件清单中存入的是文件的相对路径，相对路径是指数据在存档数据包中的相对位置，属于逻辑路径。例如，版本v2中ch3.pdf文件在总文件清单中为“fb2f...71be: [v2/ch3.pdf]”，但在版本文件清单中校验和部分相同，路径部分为“[ch3.pdf]”。

当需要进行数字对象的重建时，版本文件清单中的相对路径为数字对象的文件保存结构，通过文件对应的校验和值在总文件清单中查询文件的实际保存路径以完成重建。例如，数字对象版本v1中的ch1.pdf文件在版本v2中改变保存位置放在新增的temp文件夹下保存，由于文件本身未做修改，所以无须新增文件存储和总文件清单，只需在版本文件清单中将ch1.pdf的相对路径记录为“[temp/ch1.pdf]”。区别路径将数字对象逻辑结构和实际存储结构分离，便于对仅修改数据保存位置或文件名称而未做内容改变的文件进行重建，从而提高文件结构频繁变化需求下的数字对象保存能力。

新增版本时，保存管理文件无须改变版本内容部分的原有内容，只需添加新的版本内容信息，并对应更新文件清单部分的属性条目，使文件储存结构与保存管理文件内外逻辑统一，增强文件可读性并减少对已存储内容的修改。同时，在保存管理文件中保存先前所有版本的内容信息可极大减少版本回退或版本重建时的开销，从而在确保版本控制的同时，有效提升效率并减少成本。

4 应用效果分析

MedPRES采用Fedora作为底层仓储，Fedora的树型存储结构可以方便地支撑保存文件存储结构的设计。按照保存管理的对应关系建立“委托方→数据类型→基于协议的保存对象→版本子文件”树型文件结构（见图4）。规则清晰的文件保存结构有利于对保存内容的管理，并为保存者提供了良好的可读性和互操作性。

图4 MedPRES中的Fedora文件结构

校验和是保存管理文件中内容寻址的基础。在校验和计算方法的选择方面，考虑到相比传统的MD5和SHA1算法，SHA256具有较高的暴力破解抵抗程度，因此选择更为复杂且安全的SHA256作为内容寻址的文件校验和计算方法。对于SHA256而言，两个拥有不同内容的文件具有相同校验和值的概率为1/2^256（约为10^77），而当文件集数量接近2^128（约为10^38），才开始出现重复的校验和[21]，满足当前系统的保存需求。

为了减少新版本创建时对于保存管理文件的读取次数，同时便于保存方查看相关信息，将部分关键信息利用Fedora在页面中进行可视化的展示，如保存管理文件对应的最新版本以及文件清单中的校验和值。MedPRES中原有的元数据文件为RDF格式，为便于解析，保存管理文件同样使用RDF格式进行实现。保存管理文件相关内容如图5所示。

```
<medpres:latestVersion>v2</medpres:latestVersion>

<medpres:manifest>
    <medpres:fileList>
        <medpres:fileValue>0f3fd44b27822c9059fb7654c76d949b56578b44e4fd5b4c99a6065e4e79016e</medpres:fileValue>
        <medpres:fileLocation>v1/content/cha1.wml.xml</medpres:fileLocation>
    </medpres:fileList>
    <medpres:fileList>
        <medpres:fileValue>86a5a09f138fc2ce0a6e7114b5206df2a808b83a21a74365701f298adbee15bd</medpres:fileValue>
        <medpres:fileLocation>v1/content/ch1.pdf</medpres:fileLocation>
    </medpres:fileList>
</medpres:manifest>

<medpres:version>
    <medpres:versionNum>v1</medpres:versionNum>
    <medpres:creattime>2021-06-11T06:21:17.839Z </medpres:creattime>
    <medpres:contain>
        <medpres:containFile>
            <medpres:containChecksum>86a5a09f138fc2ce0a6e7114b5206df2a808b83a21a74365701f298adbee15bd</medpres:containChecksum>
            <medpres:containLocation>ch1.pdf</medpres:containLocation>
        </medpres:containFile>
        <medpres:containFile>
            <medpres:containChecksum>0f3fd44b27822c9059fb7654c76d949b56578b44e4fd5b4c99a6065e4e79016e</medpres:containChecksum>
            <medpres:containLocation>ch1.wml.xml</medpres:containLocation>
        </medpres:containFile>
    </medpres:contain>
</medpres:version>

<medpres:version>
    <medpres:versionNum>v2</medpres:versionNum>
    <medpres:creattime>2021-06-11T06:21:31.494Z</medpres:creattime>
    <medpres:contain>
        <medpres:containFile>
            <medpres:containChecksum>86a5a09f138fc2ce0a6e7114b5206df2a808b83a21a74365701f298adbee15bd</medpres:containChecksum>
            <medpres:containLocation>ch1.pdf</medpres:containLocation>
        </medpres:containFile>
        <medpres:containFile>
            <medpres:containChecksum>0f3fd44b27822c9059fb7654c76d949b56578b44e4fd5b4c99a6065e4e79016e</medpres:containChecksum>
            <medpres:containLocation>des/ch1.wml.xml</medpres:containLocation>
        </medpres:containFile>
    </medpres:contain>
</medpres:version>
```

图5　保存管理文件内容详情

相较于全量版本控制的方法，基于OCFL的保存文件系统提高了多版本的数据资源存取效率。无论是获取最新版本还是重建所有版本的保存对象，全量版本控制的方法都需要在MedPRES中遍历同一协议下所有版本的保存资源并找出对应的版本或按照顺序依次重建保存内容，通过采用基于OCFL的增量版本控制，仅需调用对象根目录下的保存管理文件。对于具有V个版本的保存对象，在审核过程中重建保存对象的文件系统调用成本见表2。

表2　重建对象的完整历史的文件系统调用成本

方　法	策　略	成　本
初始的全量版本控制	遍历同一协议下所有版本的目录，解析保存管理文件，并进行排序	1*V
基于OCFL保存文件系统的增量版本控制	在对应协议下的对象根目录下，解析保存管理文件	1

可见在数据量规模不大且版本数量较少的情况下，两种方法遍历保存管理文件的时间成本差别可以忽略不计。但随着数字仓储中对象数量以及版本数量的增加，两种方法的开销差距将会越大。由此可见，基于OCFL的保存文件系统设计方案更好地满足了数字资源的审计需求。

5　结语

针对数据密集型科研环境下的数字保存需求，本文基于OCFL设计了一套支持对象存储及版本控制的保存文件系统，并在医学大数据长期保存环境中进行了实践探索。设计的保存文件系统增强了对于科研数据的保存管理能力，丰富了数字保存底层文件系统对于上层应用的支持，并有效减少了数据的重复冗余保存。现有的设计仍然存在一定的局限性，如数字对象需要以既定结构保存在数字仓储中，下一步将针对数字对象分布式存储等问题开展深入研究。

参考文献

[1] HEY T, TANSLEY S, TOLLE K. The fourth paradigm: Data-intensive scientific discovery [EB/OL]. [2021-11-25]. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/?from=http%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fcollaboration%2Ffourthparadigm%2F.

[2] 刘艳红，罗健. 数据密集型科学环境下的情报服务与发展 [J]. 图书与情报，2013（6）：105-108.

[3] 柯平，袁珍珍，胡娟. 高端交流平台需要强化国家科技知识资源建设 [J]. 数字图书馆论坛，2021（3）：17-26.

[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] 彭鑫，邓仲华. 数据密集型科研环境下的科研数据管理框架研究 [J]. 数字图书馆论坛，2017（7）：61-67.

[6] 高凡，吴振新，付鸿鹊，等. 数字资源长期保存：研究进展回顾与展望——iPRES 2019国际会议综述 [J]. 信息资源管理学报，2020，10（2）：118-127.

[7] LEE C A. Open archival information system（OAIS）reference model [J]. Encyclopedia of Library and Information Sciences，2010（3）：4020-4030.

[8] 董晓莉，张炜. 基于本体的数字资源长期保存分级存储管理研究 [J]. 图书馆学研究，2017（23）：52-58，64.

[9] 王敬，王彦兵，樊向伟. 国外科研数据知识库元数据方案的调研与分析 [J]. 大学图书情报学刊，2021，39（1）：127-134.

[10] 黄鑫. 基于服务内容的科学数据服务用户满意度研究 [D]. 武汉：武汉大学，2017.

[11] HANKINSON A，BROWER D，JEFFERIES N，et al. The Oxford common file layout：a common approach to digital preservation [J]. MDPI，2019，7（2）：1-11.

[12] KUNZE J，SCANCELLA J，ADAMS C，et al. The BagIt File Packaging Format（V1.0）[EB/OL]. [2021-11-25]. http://ftp.naist.jp/pub/IETF/RFC/pdfrfc/rfc8493.txt.pdf.

[13] 吴振新，付鸿鹊，王玉菊，等. 长期保存系统数据存储管理策略研究与应用 [J]. 图书馆杂志，2017，36（9）：75-81.

[14] 张乃帅，孙超. 北京大学图书馆长期保存系统建设与探索 [J]. 大学图书馆学报，2019，37（2）：62-66.

[15] 白如江，冷伏海. “大数据”时代科学数据整合研究 [J]. 情报理论与实践，2014，37（1）：94-99.

[16] 黄鑫，邓仲华. 数据密集型科学研究的需求分析与保障 [J]. 情报理论与实践，2017，40（2）：66-70，79.

[17] JEFFERIES N，BREDENBERG K，DAPPERT A. Aligning the eARK4ALL Archival Information Package and Oxford Common File Layout Specifications [EB/OL]. [2021-11-25]. https://ipres2019.org/static/pdf/iPres2019_paper_45.pdf.

[18] HANKINSON A，JEFFERIES N，METZ R，et al. Oxford Common File Layout Specification [EB/OL]. [2021-11-25]. https://ocfl.io/1.0/spec/.

[19] 王栋，边根庆，李睿尧. 一种基于增量存储的多副本文件版本控制方法 [J]. 物联网技术，2017，7（9）：73-75.

[20] 张莲，李京，刘炜清. 云同步系统中采用增量存储的版本控制技术研究 [J]. 小型微型计算机系统，2015，36（3）：427-432.

[21] ANDERSON R. The Moab Design for Digital Object Versioning [J]. Code4Lib Journal，2013（21）：1-30.

[22] 胡佳慧，钱庆，方安，等. 医学大数据长期保存系统的设计与实现 [J]. 中华医学图书情报杂志，2019，28（9）：18-25.

作者简介

姚宽达，男，1993年生，硕士，研究实习员，研究方向：医学数据分析与知识发现、医学数字资源长期保存。
方安，男，1976年生，博士，研究馆员，研究方向：医学知识组织与数字图书馆。
杨晨柳，女，1991年生，硕士，助理研究员，研究方向：医学信息安全管理、医学数字资源长期保存。
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，通信作者，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，E-mail：hu.jiahui@imicams.ac.cn。

Design of Digital Resource Preservation File System Based on OCFL

YAO KuanDa FANG An YANG ChenLiu WANG Lei HU JiaHui
(Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, P. R. China)

Abstract: Aiming at the long-term preservation requirements of scientific research data in data-intensive scientific research environment, this paper designs a digital resource preservation file system that supports object storage and version control based on the Oxford Common File Layout method. The application implementation and effect analysis in MedPRES show the effectiveness of the designed system.

Keywords: Digital Preservation; OCFL; Data-intensive Scientific Environment; File Storage; Version Control

（收稿日期：2021-11-20）