

博物馆藏品数字资源的长期保存 系统建设

刘芳¹ 谢靖²

(1. 中国国家博物馆, 北京 100006; 2. 中国科学院文献情报中心, 北京 100190)

摘要: 随着博物馆数字资源的不断丰富, 以藏品为核心的数字资源长期保存受到越来越多的重视。本文首先研究国际敦煌、欧洲数字文化遗产平台和中国记忆3个长期保存项目的特点和应用情况; 然后创建并论述了藏品数字资源长期保存的功能模型和应用模式; 最后, 针对藏品资源特点总结开展长期保存工作应注意的关键性问题, 以期得到良好的应用效果。

关键词: 博物馆; 藏品; 数字资源; 长期保存

中图分类号: G261

DOI: 10.3772/j.issn.1673-2286.2019.03.004

博物馆是保护与传承人类历史物质文化遗产的重要场所。在近年来的信息化发展过程中, 博物馆做了大量的相关工作, 逐步实现了藏品资源的数字化建设, 但仍然存在资源格式多样、存储位置分散、存储介质不统一、检索及获取困难等问题, 不利于实现数字资源的高度共享, 同时容易造成资源丢失。在未来, 如何实现相关数字资源的统一保存和长期应用, 具有积极的研究意义。20世纪90年代起, 数字资源长期保存逐渐发展为一个研究领域。近年来, 国内外相关的研究和建设项目已具有一定规模并建立了完善的标准体系, 如荷兰的数字保存联盟(the Netherlands Coalition for Digital Preservation)^[1]和美国第三方保存系统Portico^[2]等, 形成以国家为核心的、多个机构共享的技术架构和网络环境。在我国, 国际敦煌项目(International Dunhuang Project)^[3]、中国记忆^[4]、国家数字科技文献长期保存示范体系(National Digital Preservation Program)^[5]等在历史文化和科技文献资源长期保存方面, 取得了丰富的研究成果和实践经验。本文将在研究数字资源长期保存相关技术体系、策略及案例的基础上, 结合博物馆藏品资源数字化建设的特点, 提出长期保存系统建设的初步设计思路。

1 典型案例分析

1.1 国际敦煌项目

国际敦煌项目^[3]成立于1994年, 在英国、中国、俄罗斯、日本、德国、法国和韩国设有研究中心, 共同开展敦煌西域和丝绸之路东段的文献保护、编录和数字化工作。资源类型包括绘画、手工艺品、纺织品、手稿、历史照片和地图等。截至2018年9月19日, 数据库中的影像数量达到520 681件。

(1) 资源组织。作为国际合作项目, 国际敦煌可在线提供英、中、俄、法、日、德、韩7种版本的数据, 涵盖藏品的馆藏号、出土地、收藏史、材质、尺寸、装帧、主体、文种、保存与修复情况等信息, 其中近90%的数据提供检索服务, 可在线浏览并下载藏品的高清影像, 若用于商业出版则须获得相应馆藏机构授权。此外, 项目接受资源贡献, 收集整理了敦煌学、丝绸之路相关的文献资源与藏品数据建立关联, 便于学者进行研究。

(2) 技术特点。首先, 项目建立了统一的元数据及图像命名标准。选择国际通用的元数据作为标准集, 对文字和影像格式做了严格的规范。使用XML语言和统一的DTD格式实现数据描述及交换, 所有数据均以结构化

形式存储在数据库中。其次,项目创建了统一的业务操作流程。在全球所有的成员机构设有藏品编目及数字化研究中心,对当地的成员机构进行业务培训。成员机构须按照项目组制定的工作流程完成藏品的选择与保护、数据创建、影像处理、质量检查等工作。最后,项目建立了数据管理规范及分布式数据库系统。成员机构的数据统一托管在本国研究中心的服务器上。每个成员机构拥有对本机构数据的读写权限及对其他机构数据的只读访问权限。数据的更改和添加可同步到其他服务器上。

1.2 欧洲数字文化遗产平台

欧洲数字文化遗产平台^[6]由EUROPEANA基金会于2008年组织在线开放,致力于将整个欧洲的博物馆、画廊、图书馆和档案馆的馆藏进行数字化处理并整合到平台上,便于世界各地的学者交流学习。截至2018年9月26日,可在线检索的欧洲艺术品、文物、书籍、视频和声音数据达58 246 083条,同时创建了5个主题馆藏和30多个展览,提供资源的机构超过3 500家。

(1) 资源组织。欧洲数字文化遗产平台在数字资源收集与展示方面具有更强的灵活性。首先,成员机构无须修改已有的数据结构,即可将数据资源提交到平台上共享。以德国为例,可将本国相关文化遗产机构的数据转化为LIDO格式并集成到德国数字图书馆平台(Deutsche Digitale Bibliothek, DDB)上,再由DDB将其数据统一传输至欧洲数字文化遗产平台。根据双方确定的语义转化规则,将展示元数据转化为DDB-view格式,索引元数据转化为EDM(Europeana Data Model)格式^[7]。其次,能够最大限度地各种分散、异构和跨领域资源间建立数据关联关系。如正在开发的V4Design项目可实现建筑物的3D数据建模,并与描述性数据建立关联。

(2) 技术特点。首先,欧洲数字文化遗产平台在资源的数据结构方面做了较大创新。采用自创的EDM作为数据结构标准^[6]。EDM模型参考了CIDCO的语义模型结构,定义了数据关联规则用于识别资源之间隐藏的关联关系,且与DC(Dublin Core,都柏林核心元数据)、ORE(Object Reuse and Exchange,对象复用与交换)、RDF(Resource Description Framework,资源描述框架)、FOAF(Friend-of-a-Friend, RDF词汇表)、SKOS(Simple Knowledge Organization System,简单知识组织系统)、OWL(Web Ontology Language,网络本体语言)等数据结构具有良好的兼

容性和互通性。这使得项目能够尽可能完善地保存资源数据的各种描述信息。其次,创建了较为精确的相似度算法,通过算法匹配在地理位置、时间序列、事件、主题内容、形状等情境中为资源建立关联,丰富藏品数据的情境信息。最后,创建了Europeana数据许可框架,提高与其他应用程序和服务的数据互操作性,为分布式数据库系统的建立创造了条件。

1.3 中国记忆

中国记忆项目由国家图书馆发起,以口述影像作为主要载体,照片、书信、日记、实物、文献资料为辅助载体,记录中国现代及当代“记忆”的数字资源不完整^[8]。项目资源以专题库的形式进行整合及展示,涉及的主题包括非物质文化遗产、重大历史事件和人物、图书馆记忆,数据总量在40TB以上。

(1) 资源组织。中国记忆项目收集并整理了大量的人物口述视频,与其他描述性信息共同作为历史资料划分到不同的子专题下,实现关联展示。以东北抗联专题为例,在其下分为抗联简介、老战士口述、历史照片、文献目录等子专题,这些子专题按资源类型进行分类。目前提供“我们的文字”“蚕丝织绣”“中国当代音乐家”“中国年画”等6个资源专题的在线展示和检索^[4]。未来,中国记忆项目会加强与地方的合作,利用已有的资源标准规范,收集并展示各地区、民族、行业的历史文化和实践经验。

(2) 技术特点。首先,制定了多媒体资源的描述规范^[9]。为提高视频资源的检索准确性,除视频编目标准、馆藏和版权等描述信息外,同时将人物音频转换为文字,然后以文字为基础实现对视频的内容描述和关键词标引(如口述事件发生的空间和时间范围以及视频内容描述等),为实现基于文本的关键视频片段精准查找提供了条件。其次,在视频资源与其他类型信息间建立数据关联关系。在数据存储时采用基于主题的层级结构,在每个子专题下又分别采用不同的资源组织形式并制定了严格的编码规则。

2 藏品数字资源长期保存系统建设

2.1 藏品数字资源特征分析

藏品数字资源具有5个特征:①以藏品的二维影

像、三维影像、视频等多媒体资源作为主要存储类型，而传统图书馆、档案馆以文献记录作为主要存储类型；②不同类型的多媒体资源由于采集设备和来源渠道不同，从而存在多种编码方式；③因长期保存需要，数字资源具有采集精度高、数据容量大、资源消耗多的特点；④与知识描述紧密关联，需要对数字媒体与相关文本知识关联标注；⑤为适应展厅、网站、手机终端等多种展示要求，数字资源须支持不同比例与分辨率的压缩，满足不同环境场景下的网络传输与设备展示。根据藏品数字资源的5个特征，在长期保存系统的设计和规划中，需要同时满足博物馆资源的长期保存职能和展览职能。

2.2 藏品数字资源保存系统的模型

数字资源长期保存，可以定义为一系列对数字信息进行持续管理和维护的活动，目标是确保数字信息的长期存活，保证数字信息真实可信，能够被未来的使用者所理解和应用^[10]。基于对国内外主要数字资源长期保存体系的研究和分析，其架构体系按照职能的不同，可以分为数据管理、数据摄入、数据保存、数据访问和分布式数据存储5部分。5个模块的功能划分和相互关系如图1所示。

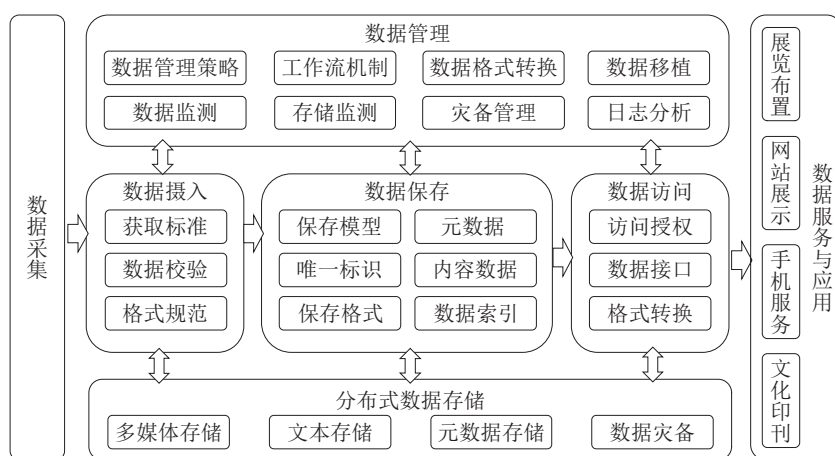


图1 数字长期保存系统的功能模型^[10]

数据管理功能模块主要用于实现资源规划、工作流程、技术策略、大规模数据迁移策略、数据存储策略、数据获取和访问策略，以及知识产权保护策略等的制定和确认工作，从技术和管理层面对整个系统的运行做出设计，并实时监测系统的运行状态。数据长期保存主要面向多个机构在某个行业、领域或专题的数据，其目的在于实现对数据的整合、永久保存和共享，因此涉及大量的管理、规划和安全保障工作，实现系统的可信赖性。

数据结构主要用于管理数字内容、元数据，及二者之间的相互关系，以便于实现对数据的导入、导出、查询和访问等功能。数据结构包括描述信息和管理信息两种资源类型，并需要满足8项功能：①数据结构须具有可扩展性，可根据保存目标及规模进行调整；②能够为数字资源及其元数据提供唯一标识符；③能够抽取和创建元数据并进行管理；④能够将资源呈现给用户；⑤需要具备强大的数据索引及检索功能；⑥对所用技术进行监测，及时发现并替换过时的技术；⑦能够记录所有元数据的历史变更信息；⑧定期对数据进行审计。

数据获取功能主要用于与资源第三方协商确定获取数据的内容、规范与版权条例后，应用数据封装、唯一标识、完整性校验和格式转换等技术手段将资源获取到长期保存系统进行存储的过程。为降低该功能模块的复杂性，一般会要求对同一机构的同类型数据采用同一种数据结构和规范进行获取。在数据进入长期保存系统后，要对数据质量进行检测，包括数据完整性、格式统一性、数据有效性、数据可读性和病毒检测等；在检测通过后，数据会进入存储模块，同时向资源提供者发出确认通知并生成系统日志。

数据存储功能主要用于解决数据的安全问题，通过数据存储软件与硬件技术相配合的方式，创建数据安全体系，增强系统对各种故障、意外事故和灾难的抵御能力。目前常用的数据存储技术包括网络存储、分层存储、虚拟化和云存储等类型。对于重要的数据资源，应实现远程备份。由于数据长期保存系统通常会涉及大体量的数据存储问题，对存储空间的容量、存储安全性、错误检测和数据读取速度具有较高要求，导致存储

设备的投入成本较高。

数据访问功能模块主要用于接收用户的资源访问请求,然后根据用户的资源访问权限和数据管理控制要求,将生成的响应数据传递给用户的过程。其中,资源定位、用户认证授权和数据开放协议等技术是数据访问模块的核心问题。应当在保证数据和系统安全性的前提下,提高系统的易用性和可用性。

2.3 藏品长期保存系统的应用模式

博物馆藏品及数字资源长期分散保存于各博物馆

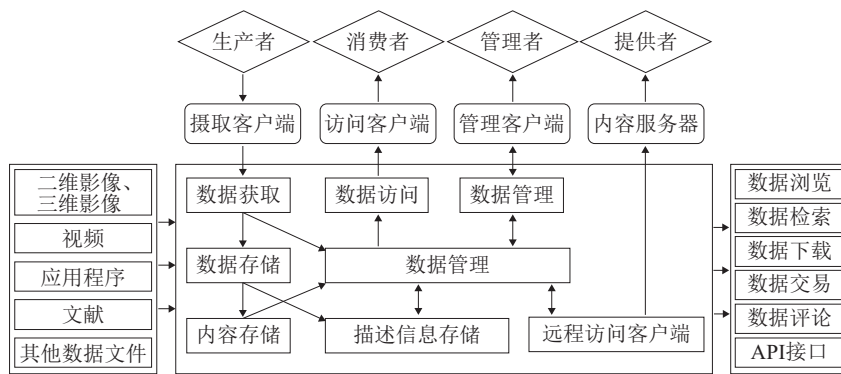


图2 长期保存系统的应用模式^[12]

在该应用模式中,各机构共同制定统一的数据管理规划,并按照规划标准将整理过的本单位资源数据保存到系统中实现资源共享。资源共享包括两种形式:对于资源体量较大、资金和人力较雄厚的机构可在本单位建立数据环境,为其他机构提供数据远程访问接口;对于资源体量较小、资金和人力较匮乏的机构,可通过摄取客户端将本单位数据存储到数据中心的服务器中。参与机构和获得授权的消费者可共享并浏览长期保存系统中的藏品数字资源,并依据制度通过系统的版权保护模块付费获得使用版权,由中心系统管理者负责数据和系统的维护工作,并提供技术支持。在该应用模式下,机构依然拥有对各自数据的控制和管理权,不会受到数据中心的强制性控制,同时可以最大限度地共享资源,提高资源的利用率。

目前,国际上关于长期保存体系建设已经形成一些较为成熟的标准体系和技术产品。2003年,美国空间数据系统咨询委员会发布的OAIS (Open Archival Information System)^[13]标准,详细描述了系统建设在技术模型、信息结构、功能结构和支撑环境等方面的要求。此外,基于LOCKSS^[14]、DAITSS^[15]、Fedora Repository^[16]等技术开

和科研院所中,这导致无论从时间、材质、器物类别、出土地等任何一个分类维度都无法获得完整的数据,给历史学、考古学、器物学的研究者和相关爱好者获取有用信息带来较大阻力。2012年,国务院启动第一次全国可移动文物普查工作,以藏品名称、藏品编号、文物类别等14个指标项作为核心元素,收集和整理了全国相关单位的藏品信息^[11]。但各研究机构还有很多在长期工作中积累的藏品信息,因为各种原因没有得到有效的收集和利用。这些问题可以通过建立分布式的藏品数字资源保存系统得到改善,应用模式如图2所示。

发的数字保存网络也值得借鉴。

3 现存问题及对策分析

3.1 加强机构、领域间的资源合作与共享

数字资源的长期保存是一个综合性的工程,目的是集合多个机构的资源和研究成果,实现某个领域资源的合作与共享。此外,不同机构存在各自的优势和不足。以综合性博物馆为例,涉及藏品资源类型丰富但并不全面,要建立整套数据标准存在较大难度且单一建立的标准难以实现未来藏品领域信息的整合。这意味着,在藏品资源长期保存建设中,应深入实现各博物馆之间的交流与合作,打破机构、区域间的地理界限,共同制定数据长期保存策略。

在制定策略时,应注意3个问题。①做好数字资源保存的规划工作。应根据藏品数字资源长期保存系统建设的目的,有针对性地选择资源保存范围并在机构间达成共识。可采用“确定数据范围及基本信息—选择实验样本—制定候选方案—分析评估结果—形成保存

规划”的方式制定方案。②建立全面的数据标准体系。长期保存的标准体系,从宏观上可以划分为资源描述信息和业务流程标准两部分。其中,资源描述信息包括字符编码标准、唯一标识、数据结构、数据封装标准与格式等内容;业务流程标准涵盖数据获取、数据存储和数据访问3个主要功能模块的业务流程标准。③协商保存数字资源的知识产权,确保数字资源的安全访问控制。确定数字资源的所有者,依照相关知识产权的管理方法,明确版权所有者的数据提供和共享权益。

3.2 应对数字资源的复杂性

复杂性和多样性是博物馆藏品的特点。除藏品的名称、材质、形状、保存状态、保存位置等基本描述信息外,还包括藏品的影像、参考文献、修复情况、展出历史等信息,涉及文本文件、二维影像、三维影像、音频文件、视频文件等多种资源类型,每种资源类型又包括多种资源格式。此外,博物馆藏品涉及书画、器物、古籍等多种类型,在颜色、质地、体积、重量等方面存在较大差别,在数字化过程中难以统一标准。

基于以上问题,在进行藏品资源长期保存工作时,可参照“标准制定—技术工具选择—工作规范—规模化实现”的步骤和流程进行建设^[17]。①建立完善的数据格式标准。可考虑基于藏品类型进行分类,为每种类型的藏品建立统一的数据采集标准和数据检验标准,如事先确定图像的色温色调和分辨率、音频和视频资源的数码格式等。在条件允许的情况下,可考虑在藏品的多媒体资源中嵌入描述性元数据,并与藏品的描述性数据建立关联,提高数据的互操作性和可分析性。②技术工具选择。尽可能做到使用同样的数据采集工具和通用的、生命周期较长的资源格式,以简化数据获取、存储和访问的难度,延长数字资源生命周期。③确立工作规范。确定多媒体由制作、修饰、编录、审核到上传的工作规范,并编制文档。④规模化实现。加大培训力度,并确保规范的全面执行。

3.3 提高数字资源的互操作性

由于藏品资源系统的初始建设目的不同,在分类标准和著录规则上存在的区别导致大量数据以异构的形式存在。这使得数字资源在整合时,可能出现结构不匹配、内容不一致等情况,成为构建藏品数字资源长期保

存系统的一个重要风险。重新制定数据标准,虽然可以从根本上解决数据异构问题,但是需要消耗极大的人力和物力成本,不是最优的解决策略。

在进行数字资源长期保存建设时,可根据需求,选择以下4种方式提高数字资源的互操作性^[18]。①建立顶层元数据应用规范,兼容多种不同的元数据标准,保证不同标准间有基本的通用结构和元素。但这种方式难以继承已有元数据标准的描述细节和深度。②在不同数据规范之间建立映射关系。以一种灵活性和兼容性较强的数据标准作为基准,在该标准和其他标准之间分别制定一对一的元数据映射规范。这种方式准确性较高,在资源整合项目中应用广泛,但在数据结构较为复杂的情况下会增加大量工作。③基于语义的数据复用与集成。语义模型采用三元组关系构建树形数据结构实现资源描述,具有较强的延伸性和扩展性,能够有效整合多种数据标准。在数据规范确定的情况下,可以使用语义结构实现不同标准间的互操作^[19]。④在分布式数据长期保存环境中,可以通过制定数据互操作协议或使用API来实现异构数据库之间的跨库检索和资源调用,不过二者多关注于底层数据标准规范和格式的互操作性。

3.4 确保数字资源的安全性

数字资源依赖数字化技术而存在,需要特定的软件技术及硬件环境才能被存储、交换和呈现。这些特质,决定数字资源具有脆弱性并容易受到外界的干扰和威胁。如人为的操作和篡改、信息技术的快速退化、物理载体的腐化、组织机构的变化、经济的衰退和自然灾害的威胁等,都会给数字资源带来无法恢复的损失。

长期保存本质上是实现对数字资源的持续管理和维护,在工作中应该注意5个问题^[20]。①重视数字内容格式的选择。格式须具有较高的标准化程度,能够在多种环境中正确地识别、理解和呈现资源内容,具有广泛的用户支持度。②根据资源的重要性,分别确立信息等级安全保护标准、可信性认证标准、审计与认证标准、数据备份与恢复方案、应急处理方案等。确保系统的安全性、技术的可靠性和软硬件的可持续使用性。③应用访问控制技术、用户认证和授权技术等,在保证用户获得资源的真实性的同时防止非法用户的网络入侵。④加强物理访问控制。相关人员须有相应的职责和权限,建立门禁安保系统,限制不相干人员接触系统的物理设施。在出现紧急情况时,须有相应的灾难恢复

机制。⑤建立长期的发展机制。保证项目经费的持续供给,确保每个业务流程都有制度可依,不会因为组织机构及人员的变动影响系统的正常运行。

4 结语

本文着重研究了国际敦煌、欧洲数字文化遗产平台和中国记忆3个数字资源长期保存项目的特点和应用情况,基于已有研究成果设计了数字资源长期保存体系的功能模型和应用模式。经过以上研究和分析,笔者认为,加强机构领域间的资源合作与共享,应对数字资源的复杂性,提高数据资源的互操作性,确保数据的安全性是博物馆藏品数字资源长期保存的关键性问题。数字化技术的发展,对博物馆传统的工作方式造成较大的冲击。在新的形势下,数字化展示、文创产品研发、用户数据分析等新型业务模式的发展都需要有强大的数据基础出作为支撑。数字资源长期保存系统作为底层建筑的作用不容忽视,应成为一项长期工作持续开展和完善。

参考文献

- [1] The Netherlands Coalition for Digital Preservation [EB/OL]. [2018-09-18]. <http://www.ncdd.nl/en/>.
- [2] PORTICO [EB/OL]. [2018-09-18]. <https://www.portico.org/>.
- [3] The International Dunhuang Project: The Silk Road Online [EB/OL]. [2018-09-18]. <http://idp.bl.uk/>.
- [4] 中国记忆 [EB/OL]. [2018-09-18]. <http://www.nlc.cn/cmptest/>.
- [5] 国家科技数字文献资源长期保存体系 [EB/OL]. [2018-09-18]. <http://www.ndpp.ac.cn/>.
- [6] Europeana [EB/OL]. [2018-09-28]. <https://www.europeana.eu/portal/en>.
- [7] DOERR M, GRADMAN S, HENNICKE S. The Europeana Data Model [C] // World Library and Information Congress: General Conference and Assembly, Gothenburg: August 10-15, 2010.
- [8] 田苗, 汤更生. 中国记忆项目的构想与实践 [J]. 国家图书馆学刊, 2015, 24 (1): 3-9.
- [9] 廖永霞, 韩尉. 中国记忆项目资源组织初探 [J]. 国家图书馆学刊, 2015, 24 (1): 17-27.
- [10] 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: 7-8.
- [11] 国家文物局第一次全国可移动文物普查工作办公室. 第一次全国可移动文物普查工作手册 [M]. 北京: 文物出版社, 2014.
- [12] UC Libraries Preservation Repository [EB/OL]. [2019-02-28]. http://www.cdlib.org/services/uc3/docs/DPRDesign_Final_Dev.pdf.
- [13] Consultative Committee for Space Data Systems (CCSDS), CCSDS 650.0-B-1, Reference Model for an Open Archival Information System (OAIS), January 2002 [EB/OL]. [2019-02-19]. <http://public.ccsds.org/publication/archive/650x0bls.pdf>.
- [14] LOCKSS [EB/OL]. [2019-20-19]. <http://www.lockss.org/>.
- [15] MODS [EB/OL]. [2019-02-19]. <http://www.loc.gov/standards/mods/>.
- [16] Fedora [EB/OL]. [2019-02-19]. <http://fedorarepository.org/>.
- [17] 寇晶晶, 吴振新. 国外图像资源长期保存实践研究进展 [J]. 图书情报工作, 2017, 61 (23): 138-144.
- [18] 杨蕾, 李金芮. 国外公共数字文化资源整合元数据互操作方式研究 [J]. 图书与情报, 2015 (1): 15-21.
- [19] CHAN L M, ZENG M L. Metadata Interoperability and Standardization—a study of methodology part I [J]. D-Lib Magazine, 2014, 12 (6): 1-19.
- [20] 郑建成, 张晓林, 赵燕, 等. 数字文献资源长期保存可持续支撑机制研究 [J]. 现代图书情报技术, 2016 (12): 1-8.

作者简介

刘芳, 女, 1985年生, 硕士, 馆员, 研究方向: 博物馆数字化建设、数据组织、数据存储与服务, E-mail: liufang@chnmuseum.cn。
谢靖, 男, 1983年生, 硕士, 副研究馆员, 研究方向: 大数据计算、数据组织、知识图谱、用户画像。

Digital Preservation and Records for Collection Resources of Museum

LIU Fang¹ XIE Jing²

(1. National Museum of China, Beijing 100006, China; 2. National Science Library, Chinese Academy of Science, Beijing 100190, China)

Abstract: With the continuous enrichment of digital resources in museums, more and more attention has been paid to the long-term preservation of digital resources of collections. Firstly, this paper studies the characteristics and application of the long-term preservation projects such as the International Dunhuang Project, EUROPEANA and China Memory. Then it discusses the function architecture and application model of long-term preservation of digital resources. Finally, the key issues that should be paid attention to in the long-term preservation work are summarized for the characteristics of the collection resources, in order to obtain good application results.

Keywords: Museum; Collection; Digital Resources; Preservation

(收稿日期: 2019-02-21)