长期保存元数据在文物数字化保护项目中的应用

姜爱蓉 程变爱 郑小惠 姚飞
(清华大学图书馆, 北京 100084)

摘要: 保存元数据在数字资源长期保存过程中起着决定性作用。本文梳理分析PREMIS保存元数据的数据模型和数据字典, 提出制定我国文物数字化保护保存元数据标准规范的思路, 最后给出参考采用PREMIS 3.0并做本地化修改的文物保存元数据方案。

关键词: 保存元数据; PREMIS; 数据字典; 文物数字化保护
中图分类号: G250.76
DOI: 10.3772/j.issn.1673-2286.2020.06.001

随着信息技术的迅猛发展，数字资源成为人们在日常学习、工作与生活中获取信息的主要来源。随之而来，数字资源长期保存和长效利用的问题日益凸显，成为国际上多个领域关注的战略焦点。20世纪90年代初，欧美相继启动多个数字信息长期保存合作项目。经过20多年的发展，在保存策略、标准规范、实践应用等方面取得了丰硕成果并积累了丰富经验。由OCLC/RLG联合研究发布的PREMIS保存元数据是其中获得高度认同并影响深远的研究成果之一，在国际范围得到广泛应用。本文梳理分析PREMIS保存元数据的数据模型和数据字典并介绍其在我国文物数字化保护项目中的应用。

1 保存元数据基本概念

保存元数据被定义为“在仓储系统中对数字保存过程提供支持的信息”[1]。仓储的作用是保存数字对象，保存元数据的作用是记录数字对象本身在产生和保存过程采用的技术特征、方法与工具；记录数字对象的起源环境、保护条件、迁移方法、保存行为的责权；记录数字对象执行的所有行为的历史等信息，为数字对象的长期保存“建档存照”。即使数字对象的权利人、保管者、技术手段、法律限制和用户等全部发生变化，根据保存元数据的记录仍可确保数字对象的长期保存和访问。保存元数据为数字保存的特殊需求提供支持，是保证数字对象可获得性、可识别性、可生存能力、可表现能力、可理解能力，以及完整性和真实性的必要信息。

数字对象在保存过程中涉及多个实体（对象、权利、事件、代理、进程等），对实体、实体属性和实体之间关系的描述构成保存元数据。

保存元数据涵盖仓储系统需要记录的、必要的元数据，以确保随着时间流逝，人们在未来仍有能力利用数字对象的全部价值。保存元数据主要包含以下信息[2]。

（1）出处（provenance）。描述数字对象保管历史，记录数字对象的创建、迁移或仿真的策略，在不同阶段对数字对象所采取的行为信息、权利变化信息、验证数字对象真实性和完整性的信息，确保被保存对象没有以未记录方式有意或无意地更改。

（2）权利管理（rights management）。描述授权仓储机构保存或访问数字对象的权利信息。记录与数字保存处理相关的知识产权的属性，以及授予这种权利的依据，如法规、执照、版权等。

（3）技术和环境（technical & interpretative

environment）。描述访问、呈现和利用数字对象时所需技术和环境的信息。其不仅包括数字对象的文件格式、软硬件和系统等描述，而且包括相关“知识”信息以帮助用户在未来可以理解数字对象。如对象是一组气候观测序列的记录，技术描述信息不仅包括该序列的记录结构和每个字段含义的数据字典，还可能包括记录观测值的仪器和校准仪器的信息。

2 PREMIS数据模型与数据字典概要

2.1 PREMIS数据模型

PREMIS基于OAIS参考模型，提出了一个面向数字保存过程实体的数据模型（见图1），定义了4种主要实体（对象、权利、事件和代理）和各实体之间的关系。

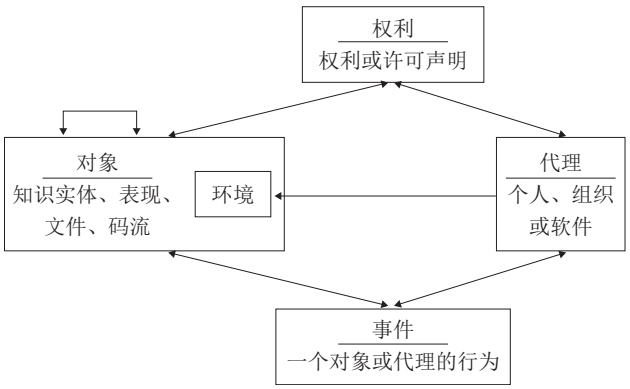

图1 PREMIS数据模型[2]

PREMIS数据模型的实体定义如下。

（1）对象实体（object entity）。以数字形式存在的离散信息单元，可以是知识实体、表现、文件或码流的形式；也可以是环境，即以某种方式（如呈现或执行）支持数字对象的技术（软件或硬件）。知识实体由连贯的内容构成，如一本书、一张图片或一个数据库等。

（2）事件实体（event entity）。涉及或影响至少一个对象或代理的行为。

（3）代理实体（agent entity）。保存事件中对某个对象实施某项操作的个人、组织或软件程序/系统，在对象的生命周期内与事件相关或与对象的权利相关。

（4）权利实体（rights entity）。属于对象与/或代理的一种或多种权利或许可的声明。

PREMIS数据模型基于OAIS概念模型的存档信息包建立，着重映射数字对象的表现信息及保存描述信息，可以理解为OAIS概念模型到可执行语义单元的翻译框架[3-4]。

2.2 PREMIS数据字典

从保存元数据的元素具有语义性、易于机器语言（XMLschema）表达和利于数据交换等角度出发，PREMIS采用语义单元（semantic unit）描述实体、实体属性和实体关系，不采用常规的元数据元素（metadata element）。语义单元的集合构成PREMIS数据字典，语义单元的结构层次可以是包括一组语义单元的容器或一个独立的语义单元，在容器层次下的语义单元仍可以是包括一组语义单元的容器或一个语义单元，这种层次化结构使得保存元数据方案更加灵活和更利于元素扩展。包含一组语义单元的容器不能被赋值，末梢的语义单元可直接赋值。语义单元的适用性、必备性和重复性在PREMIS数据字典中作出了规定。

2015年6月发布的PREMIS3.0数据字典共包括197个语义单元（含容器和独立语义单元）[5]，其中对象实体语义单元90个、事件实体语义单元21个、代理实体语义单元18个、权利实体语义单元68个，其中一级语义单元共33个（见图2）。

对象实体语义单元	事件实体语义单元	代理实体语义单元	权利实体语义单元
1.1对象标识符（objectIdentifier）	2.1 事件标识符（eventIdentifier）	3.1 代理标识符（agentIdentifier）	4.1 权利声明（rightsStatement）
1.2 对象类型（objectCategory）	2.2 事件类型（eventType）	3.2 代理名称（agentName）	4.2 权利扩展（rightExtension）
1.3 保存级别（preservationLevel）	2.3 事件日期（eventDateTime）	3.3 代理类型（agentType）	
1.4 重要属性（significantProperties）	2.4 事件详情（eventDetail）	3.4 代理版本（agentVersion）	
1.5 对象特征（objectCharacteristics）	2.5 事件结果信息（eventOutcomeInformation）	3.5 代理说明（agentNote）	
1.6 原始文件名称（originalName）	2.6 链接代理标识符（linkingAgentIdentifier）	3.6 代理扩展（agentExtension）	
1.7 存储（storage）	2.7 链接对象标识符（linkingObjectIdentifier）	3.7 链接事件标识符（linkingEventIdentifier）	
1.8 签名信息（signatureInformation）		3.8 链接权利标识符（linkingRightsIdentifier）	
1.9 环境功能（environmentFunction）		3.9 链接环境标识符（linkingEnvironmentIdentifier）	
1.10环境命名（environmentDesignation）			
1.11环境注册（environmentRegistry）			
1.12环境扩展（environmentExtension）			
1.13关系（relationship）			
1.14 链接事件标识符（linkingEventIdentifier）			
1.15 链接权利标识符（linkingRightsIdentifier）			

图2 PREMIS3.0数据字典一级语义单元

尽管PREMIS3.0数据字典包括197个语义单元，但在其应用指南部分指出：数据字典仅是一个“大多数保存仓储可能需要知道以支持数字保存事件”的核心集，建议实施中制定保存元数据方案时，不同方案、不同系统之间，甚至一个方案的不同实施实例之间，需要因地制宜，具体问题具体对待。

3 保存元数据实施要点

3.1 完整的保存元数据方案

一个完整的保存元数据方案包括PREMIS“核心”保存元数据、描述元数据、结构元数据、对象格式元数据、权利元数据等。

PREMIS数据字典定义了一组可实施的“核心”保存元数据，可理解为保存元数据的一个子集（从完整的保存元数据广义层面上）。如图3所示，中间灰色椭圆形图形为保存元数据；中间深灰色部分图形为PREMIS。

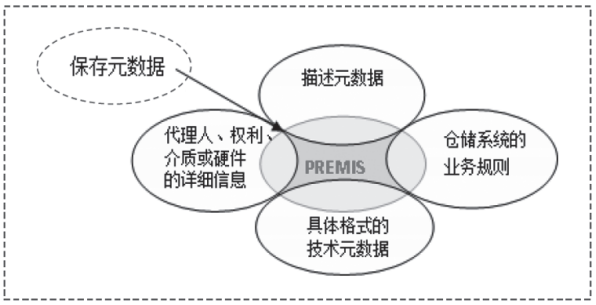

图3 PREMIS与完整保存元数据的关系[6]

PREMIS不考虑用于发现和获取的描述元数据[7]，也不定义具体格式元数据，而是专注定义大多数的保存系统在大多数情况下执行保存功能需要定义的元数据。如果某项元数据有其他机构制定且已存在标准定义，即使它与保存行为密切相关，也被排除在PREMIS元数据之外。

实施中建议：①描述元数据、结构元数据、权利元数据推荐引用已存在的元数据方案，这些类型的元数据应用广泛；②对象格式元数据记录数字对象的重要特征，是保存系统操作的重要依据。推荐参考成功案例扩展，如美国国会图书馆研制的MIX标准《数据字典——静态数字图像技术元数据》（ANSI-NISO Z39.87）[8]、音频数据词典（Audio Data Dictionary）[9]、视频数据词典（Video Data Dictionary）[10]，澳大利亚国家图书馆在APSR保存项目中对后两个数据词典做了进一步扩展[11]。

3.2 保存元数据取值的规范化、自动化

大多数保存系统面对海量数据处理。保存元数据实施的一个关键因素是元数据值是否规范化并自动抽取。为此，保存系统需要：①预先定义一些命名域和受控词表，并引用已有的标准规范；②开发自动抽取软件及工具，从待保存的海量数据中获取元数据值，这也是保存元数据有别于其他元数据的一个最大特征。

实施中取值规范化建议开发相关的唯一标识符命名域，定义若干受控词表，采用规范化的日期、时间格式和地域表达格式等。以定义受控词表为例，给出取值建议（并非穷尽枚举）：①保存级别（preservation LevelValue），建议取值完全保存、比特级保存；②格式名称（formatName），建议取值Text/sgml、image/tiff/geotiff、Adobe PDF、unknown；③内容位置（contentLocation），建议取值URI、hdl、NTFS、EXT3；④存储载体（storageMedium），建议取值磁带、光盘、本地硬盘、移动硬盘、云存储；⑤事件类型（eventType），建议取值采集、压缩、创建、解压缩、解密、删除、数字签名识别、传播、固定性检查、摄取、迁移、复制、确认、查毒；⑥代理类型（agentType），建议取值人、机构、软件；⑦版权状态（copyrightStatus），建议取值保留版权、公众领域、未知；⑧行为（act），建议取值复制、迁移、修改、使用、传播、删除。

3.3 元数据值自动抽取工具

在保存元数据的实施中，通常数字对象的有关信息（包括元数据）未被显式记录。元数据值的获取主要有两个途径：一是保存系统本身的记录、策略和文档；二是从数字对象中抽取。前者大多数情况下是系统批量赋值，后者基本上是系统自动抽取数值。从数字对象抽取元数据数值的过程包括：①预处理数字文档，剔除在格式、内容等方面存在问题或严重缺失的文档；②由元数据抽取模块批量处理数字对象，获取符合规范的元数据数值并保存；③建立保存系统与抽取获得元数据值（语义单元）的映射关系，批量摄入保存系统。

国外的长期保存系统建设已积累大量经验，开发了若干元数据自动抽取工具[12]。如英国国家档案馆的DROID文件格式识别工具[13-14]、新西兰国家图书馆的NLNZMetadata Extractor软件[15]、美国JSTOR和哈佛大学联合研制的JHOVE[16]，以及某些商业软件[17]，这些工具可以不同程度地识别现存的文件格式、版本信息等。

4 我国文物数字化保护项目中的保存元数据方案

4.1 研究背景和思路

文物数字化自20世纪90年代中期兴起以来，在世界范围内迅猛发展。越来越多的博物馆和文物收藏机构通过数字化手段再现文物古迹，提供文物展览服务和学术研究，随之而来的文物数字化保存成为一项迫在眉睫的任务。我国文博领域在“数字敦煌”“数字故宫”等文物数字化的进程中，不仅重视文物的数字化管理和数字化利用，也非常重视文物数字化的长期保存。2014年由国家文物局组织并获国家科技部“国家科技支撑计划”资金支持的“文物数字化保护标准体系及关键标准研究与示范”项目（2014BAK07B00）正式批准，由文博、图书馆和计算机等领域约30多个机构分工合作开展标准体系及关键标准的研制和应用试验。

该项目包括四大类研究课题，“文物数字化保护元数据标准研究”的成果包括63个标准规范，清华大学图书馆承担的子课题《文物数字化保护保存元数据规范》和《文物数字化保护保存元数据应用指南》，主要研究保存元数据及其在文博领域的应用，本文基于这两个子课题成果基础撰写。子课题研究思路：首先，广泛调研国内外数字图书馆、文物数字化保护项目采用的保存元数据方案，对重要项目的方案进行比较分析；其次，调研国内文物界对数字化保护的需求和应用案例；最后，结合文物数字对象的主要特征对照分析PREMIS“核心”元数据在我国本地化应用的可行性。经过理论调研和应用案例分析，选择PREMIS3.0数据字典作为我国文物数字化保护保存元数据的框架和基础，研制我国文物领域保存元数据方案并探索本地化应用。

4.2 PREMIS本地化应用的理论基础

OAIS参考模型（ISO 14721）为数字保存系统提供了一个基本框架，对数字存档系统相关的环境、功能模块以及信息对象给出了一个概念化定义。OAIS在广义层面上阐述了需要采用什么类型的信息保存数字对象。这个参考模型是实际系统建立的一个起点，从概念上描述了实际系统必须的高层任务、服务以及信息需求。PREMIS在OAIS框架的基础上更接近实际系统，提供了数字保存系统中保存元数据的框架并定义了保存系统所需的“语义单元”。PREMIS数据字典对数字保存系统的元数据管理来说是一个全面且有实际指导意义的参考工具，它定义了一套可执行的核心保存元数据，同时还制定了应用指南。PREMIS数据字典的发布将保存元数据的研究从理论向实践推进了一大步，这是选用PREMIS数据字典作为我国文物数字化保护保存元数据框架的理论基础。借鉴国际上已有的保存元数据标准，不仅有利于我国文物数字化保护保存元数据在一个高起点起步，而且有利于国际范围内保存系统之间的对接和数据交换。

4.3 PREMIS本地化应用的实践基础

2005年PREMIS发布以来，在国内外各领域的保存机构引起广泛关注和踊跃实践。首先，在美国范围，国会图书馆、哈佛大学图书馆、耶鲁大学图书馆、斯坦福大学图书馆、佛罗里达图书馆自动化中心、OCLC/RLG、美国政府等机构或联盟在保存系统和保存元数据方面的工作都基于或参照了PREMIS；其次，在欧洲范围，大英图书馆、荷兰国家图书馆、德国数字信息长期存档发展合作项目Kopal、瑞典国家档案馆、法国国家图书馆、苏格兰国家档案馆、西班牙文化部等机构或联盟的研究和实践都最大限度地参照和应用了PREMIS；再次，在泛太平洋地区，澳大利亚国家图书馆、新西兰国家图书馆、日本国立图书馆也参照或借鉴了PREMIS；另外，第三方存储Portico系统、商业化Exlibris公司的长期保存系统Rosetta等也基于或参考了PREMIS；在国内，国家“十五”重点文化建设项目——国家数字图书馆工程的“长期保存元数据规范和封装规范”项目由清华大学图书馆承担。根据国家图书数字资源长期保存的需求，清华大学图书馆基于PREMIS2.1研制了《国家数字图书馆长期保存元数据规范和应用指南》《国家数字图书馆长期保存元数据封装规范和应用指南》，该项目于2012年通过验收，两份规范和应用指南已于2014年正式出版[18-19]。在该项目基础上，国家图书馆和清华大学图书馆、上海图书馆等机构联合起草，2012年文化部发布了行业技术性指导文件《图书馆数字资源长期保存元数据规范》（WH/Z 1-2012）。中国科学院文献情报中心2007年以来致力于开发“数字资源长期保存示范系统”，后续发展为“国家数字科技文献资源长期保存体系示范系统软件平

台”，并陆续部署节点建设[20]。该保存系统基于Fedora的FOXML框架，部分元数据采用了PREMIS。由此可见，在长期保存研究和实践方面位于世界前列的大型机构和图书馆基本上都参考了或完全参照PREMIS模型构建其保存元数据框架并加以实践。PREMIS已成为数字资源长期保存领域保存元数据的“事实”标准。国内外这些案例正是清华大学图书馆子课题组选择PREMIS数据字典作为我国文物数字化保存元数据框架的实践依据。

4.4 我国文物数字化保护保存元数据方案

文物数字化保护保存元数据方案参考采用PREMIS3.0并适当修改，主要包括：①整体框架参考采用PREMIS3.0数据字典；②对197个语义单元进行了大幅精简，选取保留与我国文物数字化保存相关度较大、在实际中具有可操作性的62个语义单元（见图4）；③将语义字典中部分容器语义单元修改为可直接赋值的语义单元，使语义单元更有实际取值意义；④不在各语义单元下面设置扩展容器，而是在应用指南部分给出了扩展原则；⑤鼓励在实践中根据需求对保存元数据语义单元进行取舍或扩展。

实体	语义单元		
对象实体	对象标识符	对象标识符类型	
		对象标识符值	
	对象类型		
	保存级别	保存级别值	
		保存级别指定日期	
	重要属性	重要属性类型	
		重要属性值	
	对象特征	组分级别	
		固定性	电文摘要算法
			电文摘要
		大小	
		格式	格式名称
			格式版本
		创建程序	创建程序名称
			创建程序版本
			创建日期
		限制信息	
	原始文件名称		
	存储	内容位置	
		存储载体	
	软件环境		
	关系信息		
	关联对象标识符		
	关联事件标识符		

实体	语义单元		
事件实体	事件标识符	事件标识符类型	
		事件标识符值	
	事件类型		
	事件日期		
	事件细节		
	事件结果信息		
	链接代理标识符		
	链接对象标识符		
代理实体	代理标识符	代理标识符类型	
		代理标识符值	
	代理名称		
	代理类型		
权利实体	权利声明	权利声明标识符	权利声明标识符类型
			权利声明标识符值
		版权信息	版权状态
			版权管辖区域
			版权状态颁布日期
		权利准予	行为
			限制
			授权时间 开始日期
			结束日期
		链接代理标识符	
		链接对象标识符	

图4 文物数字化保护保存元数据全部语义单元

PREMIS数据字典定义的对象、事件、代理、权利这4个部分实体的语义单元基本涵盖了数字资源长期保存过程中需要记录的完整信息，而我国文物数字化保护的对象绝大多数为自建数字资源，事件、代理和权利实体部分相对简单。根据现实需求，文物数字化保护保存元数据方案对PREMIS3.0数据字典中的事件、代理和权利实体的语义单元做了修改和精简。同理，自建数字资源的环境也相对简单，文物数字化保护保存元数据方案也对对象实体中环境部分的语义单元做了适当精简。

5 结语

长期保存元数据方案是构建保存系统的逻辑框架，也是决定保存系统的功能特征、运行模式和总体性能的关键因素。一个良好的长期数字保存策略首先体现在与之适应的保存元数据方案中，进而通过保存系统实现。在国家文物局科技司的组织下，清华大学图书馆承担完成的《文物数字化保护保存元数据规范》和《文物数字化保护保存元数据应用指南》正在形成行业标准，为文博领域的文物数字化保护项目提供参照，在我国数字资源长期保存、可互操作和可持续发展中发挥作用。

参考文献

[1] PREMIS Preservation matadata maintenance activity. Data Dictionary for Preservation Metadata Version 1.0 [EB/OL]. [2020-04-13]. http://www.oclc.org/research/projects/pmwg/premis-final.pdf.

[2] Brian Lavoie and Richard Gartner. Preservation Metadata (2nd edition). DPC technology watch Report [EB/OL]. [2020-04-13]. https://www.dpconline.org/docs/technology-watch-reports/894-dpctw13-03/file.

[3] 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.

[4] 刘建华，张智雄. 保存元数据的发展趋势研究 [J]. 图书馆杂志，2016 (6)：10-16.

[5] PREMIS Data Dictionary for Preservation Metadata, Version 3.0 [EB/OL]. [2020-04-14]. http://www.loc.gov/standards/premis/v3/index.html.

[6] Priscilla Caplan. Understanding PREMIS: an overview of the PREMIS Data Dictionary for Preservation Metadata [EB/OL]. [2020-04-14]. http://www.loc.gov/standards/premis/understanding-premis.pdf.

[7] 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.

[8] MIX: NISO Metadata for Images in XML Standard [EB/OL]. [2020-04-14]. http://www.loc.gov/standards/mix/.

[9] Audio-Visual Prototyping Project: Audio (Source) Data Dictionary [EB/OL]. [2020-04-14]. http://www.loc.gov/rr/

mopic/avprot/DD_ASMD.html.

[10] Audio-Visual Prototyping Project：Video（Source）Data Dictionary [EB/OL]. [2020-04-14]. http://www.loc.gov/rr/mopic/avprot/DD_VSMD.html.

[11] Australian Partnership for Sustainable Repositories PREMIS Requirement Statement Project Report [EB/OL]. [2020-04-14]. https://openresearch-repository.anu.edu.au/bitstream/1885/46447/4/presta.pdf.

[12] 曾苏，马建霞，张秀秀. 元数据自动抽取研究新进展 [J]. 现代图书情报技术，2008（4）：7-11.

[13] DROID（Digital Record and Object Identification）[EB/OL]. [2020-04-22]. http://digital-preservation.github.io/droid/.

[14] 王玉菊，吴振新，孔贝贝，等. DROID开源工具在长期保存系统格式识别中的应用 [J]. 现代图书情报技术，2015（1）：75-81.

[15] Implementing the PREMIS data dictionary：a survey of approaches [EB/OL]. [2020-04-22]. http://www.loc.gov/standards/premis/implementation-report-woodyard.pdf.

[16] JHOVE-JSTOR/Harvard Object Validation Environment [EB/OL]. [2019-12-20]. http://jhove.sourceforge.net/.

[17] Metadata Extraction Tool [EB/OL]. [2020-04-26]. https://sourceforge.net/projects/meta-extractor/.

[18] 姜爱蓉，杨东波，程变爱. 国家数字图书馆长期保存元数据规范与应用指南 [M]. 北京：国家图书馆出版社，2014.

[19] 郑小惠，杨东波，童庆钧. 国家数字图书馆长期保存信息包封装标准规范与应用指南 [M]. 北京：国家图书馆出版社，2014.

[20] 吴振新，付鸿鹄. 数字信息资源分布式协作保存网络构建研究 [J]. 数字图书馆论坛，2016（9）：43-48.

作者简介

姜爱蓉，女，1955年生，研究馆员，通信作者，研究方向：数字图书馆，E-mail：jiangar@tsinghua.edu.cn。
程变爱，女，1975年生，硕士，馆员，研究方向：数字图书馆、元数据。
郑小惠，女，1970年生，硕士，副研究馆员，研究方向：数字图书馆、元数据、特藏资源组织与整理。
姚飞，女，1979年生，博士，副研究馆员，研究方向：数字图书馆。

The Application of Long-term Preservation Metadata in the Digital Protection Project of Cultural Relics

JIANG AiRong CHENG BianAi ZHENG XiaoHui YAO Fei
(Tsinghua University Library, Beijing100084, China)

Abstract: Preservation metadata plays a decisive role in the long-term preservation of digital resources. This article analyzes and summarizes the data model and data dictionary for PREMIS. It puts forward the idea of formulating preservation metadata standard specifications in China's cultural relics digital protection, and finally gives a cultural relics preservation metadata scheme that references and adopts PREMIS 3.0 and makes localized modifications.

Keywords: Preservation Metadata; PREMIS; Data Dictionary; Cultural Relics Digital Protection

（收稿日期：2020-05-19）