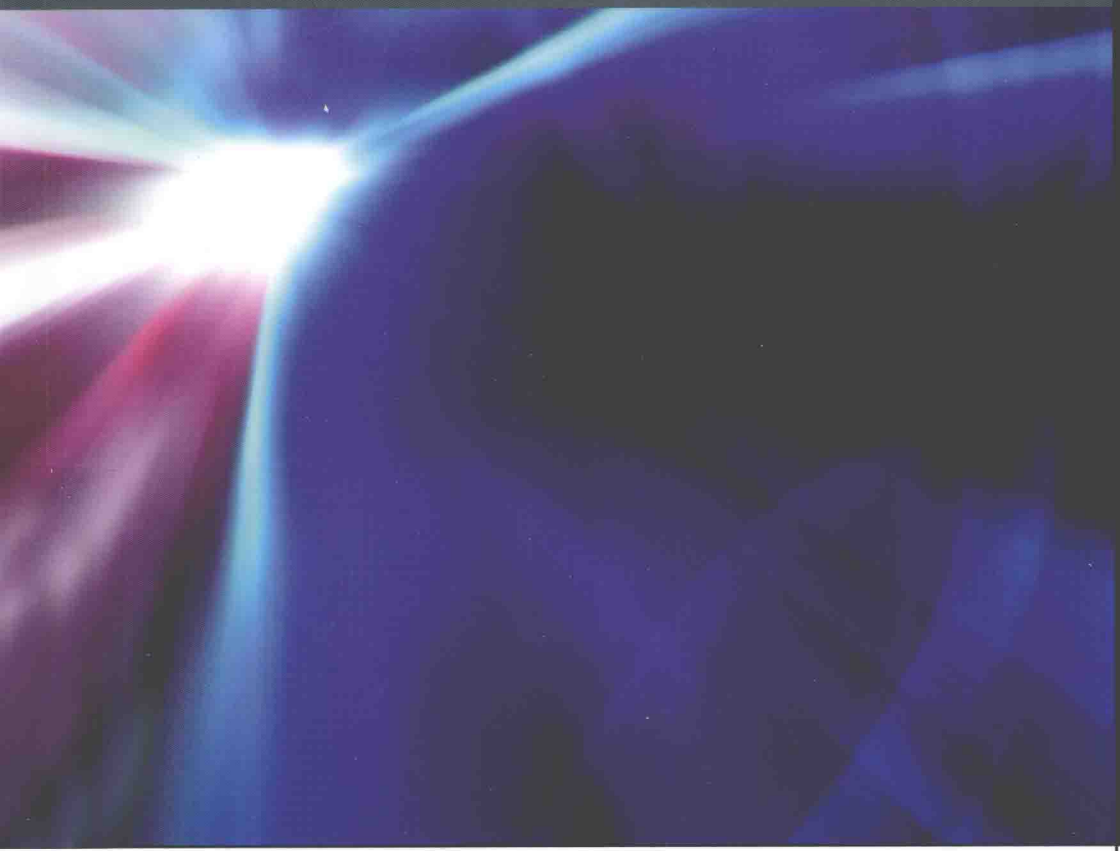
国家数字图书馆工程标准规范成果

国家数字图书馆长期保存元数据规范与应用指南

姜爱蓉　杨东波　程变爱　主编



國家圖書館出版社
National Library of China Publishing House

国家数字图书馆工程标准规范成果

国家数字图书馆长期保存元数据规范与应用指南

姜爱蓉　杨东波　程变爱　主编

國家圖書館出版社
National Library of China Publishing House

图书在版编目(CIP)数据

国家数字图书馆长期保存元数据规范与应用指南/姜爱蓉,杨东波,程变爱主编. --北京:国家图书馆出版社,2014.9

(国家数字图书馆工程标准规范成果丛书)

ISBN 978-7-5013-5444-3

Ⅰ.①国… Ⅱ.①姜… ②杨… ③程… Ⅲ.①中国国家图书馆—数字图书馆—数据技术—信息资源—规范—指南 Ⅳ.①G255.75-65

中国版本图书馆 CIP 数据核字(2014)第 187740 号

书　　名　国家数字图书馆长期保存元数据规范与应用指南

著　　者　姜爱蓉　杨东波　程变爱　主编

责任编辑　高　爽

出　　版　国家图书馆出版社(100034　北京市西城区文津街 7 号)

(原书目文献出版社　北京图书馆出版社)

发　　行　010-66114536　66126153　66151313　66175620

66121706(传真),66126156(门市部)

E-mail　btsfxb@nlc.gov.cn(邮购)

Website　www.nlcpress.com ——→投稿中心

经　　销　新华书店

印　　装　北京科信印刷有限公司

版　　次　2014 年 9 月第 1 版　2014 年 9 月第 1 次印刷

开　　本　787×1092(毫米)　1/16

印　　张　9.25

字　　数　90 千字

书　　号　ISBN 978-7-5013-5444-3

定　　价　58.00 元

《国家数字图书馆工程标准规范成果》丛书编委会

本书编委会

主编：姜爱蓉　杨东波　程变爱

编委：郑小惠　童庆钧　姚　飞　张成昱　陈　武
窦天芳　远红亮　王志庚　申晓娟　李春明
翟喜奎　赵　悦　周　晨　董晓丽　王文玲

总　　序

数字图书馆涵盖多个分布式、超大规模、可互操作的异构多媒体资源库群，面向社会公众提供全方位的知识服务。它既是知识网络，又是知识中心，同时也是一套完整的知识定位系统，并将成为未来社会公共信息的中心和枢纽。数字图书馆建设的最终目标是实现对人类知识的普遍存取，使任何群体、任何个人都能与人类知识宝库近在咫尺，随时随地从中受益，从而最终消除人们在信息获取方面的不平等。“国家图书馆二期工程暨国家数字图书馆工程”是国家“十五”重点文化建设项目，由国家图书馆主持建设，其中国家数字图书馆工程的建设内容主要包括硬件基础平台、数字图书馆应用系统和数字图书馆标准规范体系。

标准规范作为数字图书馆建设的基础，是开发利用与共建共享资源的基本保障，是保证数字图书馆的资源和服务在整个数字信息环境中可利用、可互操作和可持续发展的基础。因此，在数字图书馆建设中，应坚持标准规范建设先行的原则。国家数字图书馆标准规范体系建设围绕数字资源生命周期为主线进行构建，涉及数字图书馆建设过程中所需要的主要标准，涵盖数字内容创建、数字对象描述、数字资源组织管理、数字资源服务、数字资源长期保存五个环节，共计三十余项标准。

在国家数字图书馆标准规范建设中，国家图书馆本着合作、开放、共建的原则，引入有相关标准研制及实施经验的文献信息机构、科研机构以及企业单位承担标准规范的研制工作，这就使得国家数字图书馆标准规范的研制能够充分依托国家图书馆及各研制单位数字图书馆建设的实践与研究，使国家数字图书馆的标准规范成果具有广泛的开放性与适用性。本次出版的系列成果均经过国家图书馆验收、网上公开质询以及业界专家验收等多个验收环节，确保了标准规范成果的科学性及实用性。

目前，国内数字图书馆标准规范尚处于研究与探索性应用阶段，国家图书馆担负的职责与任务决定了我们在数字图书馆标准规范建设方面具有的责任。此次将国家数字图书馆工程标准规范研制成果付梓出版，将为其他图书馆、数字图书馆建设及相关行业数字资源建设与服务提供建设规范依据，对于推广国家数字图书馆建设成果、提高我国数字图书馆建设标准化水平、促进数字资源与服务的共建共享具有重要意义。

国家图书馆馆长　周和平

2010 年 8 月

前　言

进入到数字信息时代，信息技术高速发展，原生态数字资源和非原生态数字资源的种类和数量也越来越多。国家图书馆具有履行重点收藏和长期保存中文数字资源的职能，因此迫切需要把这些数字资源保存起来，能够让几十年，以至更长时间以后的人们看到并能够看懂。

国家图书馆的数字资源包括自建数字资源、外购数字资源、互联网采集的数字资源以及全国文化信息共享工程的数字资源等。截至 2011 年 6 月底，国家图书馆数字资源总量已达 545TB，数字资源发布总量 396.7TB，这些资源都急需有效的保存和管理。为实现国家图书馆数字资源的长期保存和有效存取，需要制定相应的长期保存标准规范。

本标准规范由国家图书馆提出，委托清华大学图书馆进行研制，为服务国家图书馆数字资源长期保存，根据国家数字图书馆工程长期保存规范招标指南、研制需求书和成交合同研制。经过广泛调研，并与国家图书馆协商，为了兼顾今后国际间的数据交换，本标准规范修改采用 OCLC 与 RLG 联合资助成立的“保存元数据：实施策略（Preservation Metadata：Implementation Strategies，PREMIS）”工作组制订的《PREMIS 保存元数据数据字典 2.1》。如要以此规范为基础发展中国长期保存元数据标准规范，需要在国家图书馆应用的基础上进行修改、补充和完善，并按照国家标准研制的正式程序组织有关单位共同研制。

长期保存元数据框架设计的基本要求是全面性、可扩展性和普适性。这就要求保存元数据的设计必须建立在一个共同的概念框架基础上，而 OAIS 不仅仅提供了这样一个概念框架，还在一个广泛的信息环境中提出了一套完整的数字资源保存系统的功能模块，并制定了信息模型，这就为我国长期保存元数据的设计提供了一个基础平台。PREMIS 数据字典提供了一个核心的保存元数据集，鉴于国家图书馆的数字资源情况，PREMIS 的这个核心元数据集是基本上适用的。考虑到国家图书馆数字资源的具体情况，实施过程可能需要进行一定的扩展，根据国家图书馆的具体需求制定了相应的应用指南。

本书提供国家数字图书馆长期保存元数据标准规范的实施建议和应用指南。为国家数字图书馆设计数字信息长期保存系统提供参考，为 PREMIS 数据模型的实施、元数据存储等提供具体的指导。对于长期保存元数据来说，其具体的实施应该是在长期保存系统中，由系统设计人员根据具体实施单位的系统需求和资源情况，将可能会用到的语义单元尽可能全地设计到

系统相关模块中，从而可以在系统运行时自动获取相应的元数据值。

一般元数据标准规范的指南基本上都是著录细则，而保存元数据与别的元数据有很大的不同，因此本书应用指南部分也完全有别于别的元数据规范的著录细则。本应用指南定位于辅助理解“国家图书馆长期保存元数据标准规范”，并对具体操作提供建议和指导，不能独立使用。

最后，特别感谢项目组成员、国家图书馆王志庚研究馆员。他对国内外的保存元数据标准和发展进行了大量研究，并将其前期成果作为本课题的研究基础，为本标准规范的研制做出不可或缺的贡献。

目　录

第一部分　国家数字图书馆长期保存元数据标准规范

1 范围

本规范修改采用《PREMIS 保存元数据数据字典 2.1》,将“保存元数据”定义为在一个保存系统中对数字资源保存过程进行支持的信息。保存元数据支持和记录数字资源保存的处理过程,此过程包括:

①创建清晰的来源记录,能够记录数字对象随时间而改变的流程;

②记录数字对象的真实性,证明没有被无法记录的方式所改变;

③记录数字对象经历的技术处理;

④描述数字对象的技术环境,包括该数字对象被呈现或利用时所需的软硬件等技术需求;

⑤描述数字对象的起源环境;

⑥指定权限管理信息,包括一定时间内限制保存系统保存和传播数字对象的权利信息。

因此,保存元数据兼有管理(包括权利和权限)元数据、技术元数据和结构元数据的功能。在保存元数据中,特别需要关注的是记录数字对象历史的来源信息和在保存系统之中数字对象之间的关系信息。

由于 PREMIS 的范围限定在对所有数字对象具备普适性的语义单元上,并不包括那些文件格式专有的元数据语义单元,所以需要根据中文数字信息资源的特点,制定专有的文件格式定义技术元数据,同时还要广泛采用国外标准的文件格式定义专有技术元数据。

本标准适用于国家图书馆数字资源长期保存方案构建及其数字资源长期保存,其他图书文献机构及其相关机构也可参照此标准开展数字资源长期保存业务。

2 规范性引用文件

本标准无规范性引用文件。列出本章是为了与其他数字资源长期保存体系标准的条款号相一致。

3 术语定义

3.1 保存元数据 preservation metadata

保存元数据是支持与数字资源长期保存相关过程的信息框架。更确切地说,它是支持数字资源长期保存过程中的可生存能力、可还原能力与可理解能力的必要信息。保存元数据能够作为保存过程中的输入信息,也可以作为相同过程的输出信息。

3.2 PREMIS 数据模型 the PREMIS data model

一种简单的数据模型，定义了数字资源保存活动中的知识、对象、事件、权利和代理五种实体。

注：知识、对象、事件、权力和代理5种实体将在第4章中详细说明。

3.3 数字对象 digital object

一个不连续的数字形式的信息单元，是 PREMIS 数据模型的核心概念，包含三个层次：文件、比特流和表现。

注：其中比特流和表现将在第4章中详细说明。

3.4 文件 file

一组0字节或多字节、可以被操作系统识别的数据。

注1：文件可以读写和复制；

注2：文件有名字和格式。

3.5 数据字典 data dictionary

依据 PREMIS 数据模型组织的，提供了数字对象、事件、代理和权力四个实体的详细描述的数据信息集合。

注：由于知识实体的元数据属于描述性元数据，应为其他既存描述元数据框架的重点研究范围，因此未被列入本数据字典。

4 PREMIS 数据模型

4.1 概述

PREMIS 工作组提出了一个简单的数据模型来组织数据字典中定义的语义单元。数据模型定义了在数字资源保存活动中尤为重要的五种实体：知识实体、对象、事件、权利和代理。数据字典中定义的每个语义单元都是数据模型实体之一的某种属性。PREMIS 数据模型，见图1。

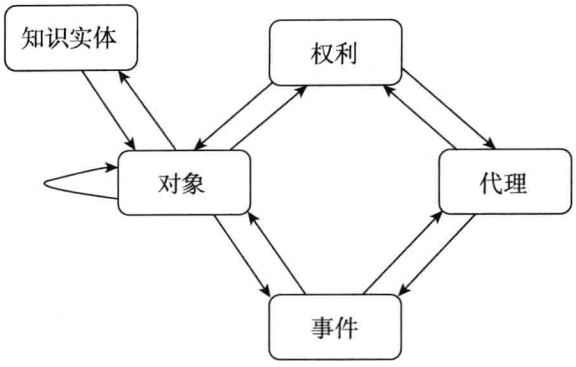


图 1　PREMIS 数据模型

在图 1 中,实体用方框表示;实体之间的关系用箭头表示。箭头的方向表示关系连接的方向,与在保存元数据中的记录一致。例如,从权利实体指向代理实体的箭头表示与权利实体相关的元数据包括一个语义单元,该单元记录了与代理的关系信息。

对象实体的箭头指向它本身,这表示数据字典中定义的语义单元支持对象之间相互关系的记录。数据模型的其他实体不支持这类关系;换言之,尽管对象可以和其他对象相关,事件却不能和其他事件相关,代理也不能和其他代理相关,如此类推。

数据模型依赖于不同实体之间的连接,使这些关系明晰。关系是指实体实例之间联系的声明。“关系”有广义和狭义之分,以多种方式表达。例如,“对象 A 的格式是 B”就可以认为是 A 和 B 之间的关系。然而,PREMIS 模型把格式 B 看作是对象 A 的一个属性。PREMIS 约定“关系”是两个或两个以上对象实体之间的联系,或者是不同类型实体之间的联系,如对象和代理之间。

4.2　对象

对象实体有三个子类型:文件、比特流和表现。

4.2.1　文件

文件是被命名的,由一组字节组成的有序序列。可以被操作系统识别。文件可以是 0 字节或更多字节,有文件格式、访问权限和文件系统特征,如大小和最后访问时间。

4.2.2　比特流

比特流是文件内连续或非连续的数据,对保存用途而言,具备有意义的共同属性。如果不加文件结构(头标等)或为遵从某种文件格式而重定比特流格式,比特流不能转换成独立的文件。

本规范中定义的比特流是指内嵌于文件的一组二进制位数。这与一般的使用有区别,理论上比特流可以包含不止一个文件。内嵌比特流文件的一个范例是包含两个图像的一个

TIFF 文件。

根据 TIFF 文件格式规范，一个 TIFF 文件必须有一个包含文件相关信息的头标。它可以包含一个或多个图像。在 PREMIS 数据模型中，每个图像都是一个比特流，可以有其属性，如标识、位置、限制信息和详细的技术元数据（如色彩空间）。

一些比特流有和文件相同的属性，一些则没有。内嵌于 TIFF 文件的图像的属性显然和文件本身不同。然而，在另一个例子中，三个 TIFF 文件可以聚合在一个大的 tar 文件里。在这种情况下，这三个 TIFF 文件也是内嵌的比特流，但它们有 TIFF 文件的所有属性。

本规范定义的比特流只包含一个内嵌的比特流，如果不加文件结构（如头标）或为遵从某种文件格式规范而重定格式，比特流不能转换成一个独立的文件。比特流的例子如 TIFF 文件内的一个图像、WAVE 文件内的音频数据，或 Microsoft Word 文件内的图像。

一些内嵌的比特流可以转换成独立的文件，尽管转换过程（如解压、解密或解码）可能会在抽取过程中对比特流进行，但不增加额外的信息。这些比特流的例子包括 tar 文件内的一个 TIFF 文件，或一个 XML 文件内一个编码的 EPS 文件。

本规范中，这些比特流定义为“文件流”，也就是内嵌于更大文件内的真正的文件。文件流有文件的所有属性，而比特流则不是这样。在数据字典中，“文件”一列应用于文件和文件流上。“比特流”一列应用于比特流（而非文件流）的子集上，依附于更严格的比特流定义。文件的位置（数据字典中的 contentLocation）通常是存储的位置；而文件流或比特流的位置通常是内嵌文件内的起始端。

比特流是文件内连续或非连续的数据，对保存用途而言，具备有意义的共同属性。

4.2.3 表现

表现是需对知识实体进行完整而合理再现的一组文件（包括结构化元数据）。例如，一篇期刊论文可以由一个 PDF 文件完成；这一个文件构成表现。另一篇期刊论文可以由一个 SGML 文件和两个图像文件组成；这三个文件构成表现。第三篇论文共 12 页，每页由 1 个 TIFF 图像表现，加上一个结构化元数据的 XML 文件，显示页面顺序；这 13 个文件构成表现。

许多保存系统的目标是随着时间推移，维护知识实体的可用版本。为了使一个知识实体显示、播放或可用，组成该知识实体至少一个版本的所有文件必须标识、存储和维护，这些文件才能被聚合起来，呈现给任意地点的用户。表现就是需要做到这些的一组文件。

对于同一个知识实体，保存系统的表现形式不止一种。例如，保存系统可以得到一个 TIFF 文件的图像（如“奔马像”），可以从 TIFF 文件创建一个派生的 JPEG2000 文件，同时保留两个文件，每个文件将构成“奔马像”的表现。

更复杂的情况，“奔马像”可能是包含该 TIFF 图像的文章片段和一个 SGML 编码的文本文件。如果保存系统创建了一个 TIFF 文件的 JPEG2000 版本，文章将有两种表现：TIFF 和 SGML

文件构成一种表现，而 JPEG2000 和 SGML 文件构成另一个表现。这些表现如何存储据实施而定。一个保存系统可以选择存储这个 SGML 文件的拷贝，在表现之间共享。保存系统也可以选择复制 SGML 文件，存储两个相同的备份。这两个表现将包括 TIFF 和 SGML 拷贝 1，以及 JPEG2000 和 SGML 拷贝 2。

不是所有保存系统都关注表现。例如，保存系统可能只保存文件对象，依靠外部代理把这些对象聚合成有用的表现。如果保存系统不管理表现，就不需要记录它们的元数据。

4.3 代理

代理是指与一个数字对象的保存事件相关联的个人、机构或软件程序。本规范没有定义代理的详细特征。只要一个保存系统能够正确识别参与某一保存行为的代理，代理的其他属性可由保存机构自行开发。

代理显然很重要，但不是数据字典的重点。数据字典只定义了一种确定代理的方式和代理类型分类（个人、组织或软件）。如果需要更多元数据，请于具体实施中自行定义。

数据模型图表中有一个从代理实体指向事件实体的箭头，但没有从代理指向对象实体的箭头。代理只能通过事件间接地影响对象。每一个事件可以有一个或多个相关对象、一个或多个相关代理。因为同一个代理可以在不同的事件扮演不同的角色，所以代理的角色是事件实体的属性，而不是代理实体的属性。

4.4 权利

权利，或称权利陈述，是与数字对象或代理相关联的一项或多项权利声明或许可。本规范仅定义了与保存行为相关的权利和许可信息，不包含那些与信息存取和信息发布相关的权利信息，即授权某一代理对某一数字对象采取某种操作（行为或限制）的许可信息。很多机构关注与知识产权和许可相关的元数据，从权利表现语言到 < indecs > 框架，然而只有很少机构提出了与数字资源保存相关的权利和许可。因此，原来数据字典中的 permissionStatement 在这一版本中改为 rightsStatement。必须指出，建议使用的 copyrightMD 和 PREMIS 权利很不一样。CopyrightMD 设计用于记录实际信息，让人对某一作品作出可靠的版权评价。PREMIS 的 rightsStatement 设计用于让保存系统决定它是否有权利以自动方式采取某项行动，附有一些声明基础的文件。

4.5 事件

事件是可被保存系统所记录的一种影响到至少一个数字对象或代理的行为。事件是记录关于数字对象行为的实体，对于本规范是很重要的，因为很多行为都将影响数字对象的保存，