国 家 数 字 图 书 馆 工 程
长 期 保 存 规 范 项 目 研 制 成 果

GC-HD090188

项目名称: 国家数字图书馆工程长期保存规范

成果名称：国家数字图书馆长期保存元数据标准规范应用指南

成果类型：应用指南

成果编号：GC-HD090188-03

成果版本：修改稿

提交时间：2012 年 4 月

研制机构：清华大学图书馆

撰 写 人：程变爱、郑小惠、童庆钧、姜爱蓉

目　录

前　言

本应用指南是国家数字图书馆工程长期保存规范项目研制成果之一。本应用指南由国家图书馆提出，委托清华大学图书馆进行研制。

应用指南的目的是为标准规范的具体实施提供指导，而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，因此本指南也完全有别于别的元数据规范的著录细则。本指南定位于辅助理解《国家图书馆长期保存元数据规范》，对国家图书馆在具体实施《规范》时提供建议和指导，不能独立使用。

本应用指南由清华大学图书馆起草，主要起草人为：程变爱、郑小惠、童庆钧、姜爱蓉。参加本指南研究工作的还有：姚飞。

1 本指南内容概述

在规范中，对于保存元数据各语义单元的定义、基本原理、著录约束、使用对象类型、必备性、重复性都做了详细的说明，还有更为详尽的具体创建、维护、使用附注说明，甚至还有一些取值范例，因此本指南中不再重复这部分，而是根据 PREMIS 数据字典的必备性定义，结合国家数字图书馆的实际情况，简单列出一些必备语义单元，以便设计人员在设计长期保存系统时参考。此外，还对超出 PREMIS 范畴的对象格式元数据进行了一定的扩展。

保存元数据实施的一个关键因素就是数据值能否由保存系统自动提供、自动进行，因此本指南中罗列出了需要保存系统开发的标识符命名域、需要保存系统定义的受控词表，以及一些语义单元取值应遵循的现有标准规范。

本指南还就 PREMIS 数据模型的实施、元数据存储、提供元数据值进行了一定的说明，以作为具体实施中的指导。另外，分析了已有长期保存系统中有关 PREMIS 语义单元的具体实现，以便国家数字图书馆在今后具体设计保存系统时借鉴。

考虑到大多数数字保存系统都需处理海量的数字信息，在数字信息长期保存的实践中，一个关键问题就是长期保存元数据的赋值是否可被自动提取并自动应用。因此，本指南介绍了现有的一些元数据自动抽取工具，作为具体系统设计中的参考。

最后，针对国家图书馆各类型的数字资源以及长期保存各环节的管理需求，分别设计了几个实用情景，以便检测《国家数字图书馆长期保存元数据规范》的科学性、合理性与实用性，同时演示长期保存元数据规范在长期保存系统中可能的记录形式。实例的取值都是示例性的，不具有真实意义。

2 保存元数据

2.1 保存元数据的定义和功能

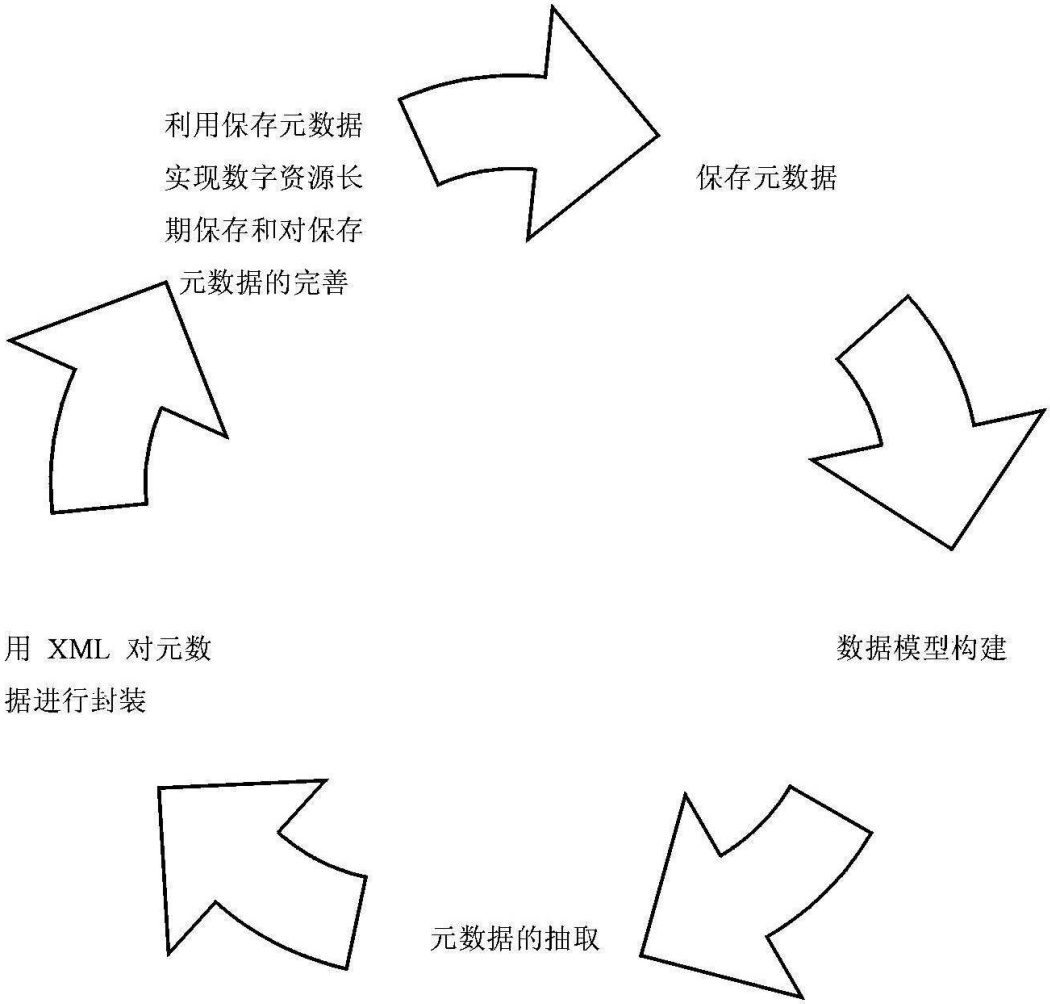
保存元数据是支持与数字资源长期保存相关过程的信息框架。更确切地说，它是支持数字资源长期保存过程中的可生存能力、可还原能力与可理解能力的必要信息。保存元数据能够作为保存过程中的输入信息，也可以作为相同过程的输出信息。

保存元数据支持和记录数字保存的处理过程，应当具有以下主要功能：

(1） 支持和证明数字保存过程的信息：创建清晰的来源记录；详细描述真实状态；记录数字对象经历的技术处理；对数字对象的技术细节进行描述；描述数字对象的起源环境；指定权限管理信息。

(2） 提供长期维护资源的信息：可生成能力（对象的比特流是完整的）；可还原能力（可以将对象转化成能够阅读或利用的格式）；可理解能力（还原的内容能被解释和理解）。

保存元数据与数字资源长期保存功能实现流程如下图所示：



2.2 完整的保存元数据方案

因 PREMIS 仅仅考虑到通用于各种格式数字对象的元数据，不包含只适于特定文件格式的技术元数据。它包含但没有详细介绍语义实体，因为“已存的标准很好地服务于描述元数据”。同时它“考虑到保存活动的权限特性，而非与存取和发布相关的权限”，所以在具体实施的过程中，实施方需要根据具体情况引用其他相关元数据方案。

一个完整的长期保存元数据方案应包括以下几个方面：

- **PREMIS“核心”保存元数据** 主要是数字对象、事件和代理实体。关于核心语义单元的说明：一方面，核心是指在任何情况下都需要的元数据；另一方面，核心是指适用于实施任何保存策略的任何系统的元数据。本应用指南中，我们认为核心语义单元即“大多数数字保存系统都可能需要获知以支持数字保存的元数据”。核心并不意味着必备，在一些例外情况下，某些语义单元是可选的。
- **描述元数据** 描述内容，包括描述数字对象上下文或意义的元数据，等价于 PREMIS 中的“语义实体”。
- **结构元数据** 保存系统中需要这种元数据来根据部分信息重构整个数字对象。它还需要

能够显示和表示一个数字对象，使用户能理解这个对象与其各部分或者与一个更大的整体间的联系。

- **对象格式元数据** 某些特定格式如图像、音频等对象的具体技术元数据信息。
- **访问权限元数据** 保存系统在提供相关记录的访问支持时必须强制执行的、用于描述限制、权限和条件的元数据。PREMIS 仅描述了保存系统需要的权限，以执行相关记录的活动。而没有详细制定访问权限管理体制，因为它涉及到不同的特定数据集所具备的不同权限管理机制，太复杂并不断地演化。国家图书馆可以开展一个访问权限管理项目，用于处理各类型数字和非数字馆藏。同时从元数据分类划分看，该部分内容属于管理元数据的范畴，故不在此指南中涉及。

2.3 必备语义单元

2.3.1 概述

PREMIS 的数据字典提供了一个“核心保存元数据语义单元集”。就国家图书馆数字资源的长期保存而言，认为语义单元如果适用都是必需的，接下来就是如何具体实施 PREMIS 的问题了。PREMIS 数据字典定义的核心保存元数据集在规范中已经做了详细介绍，本指南中不再重复。根据 PREMIS 数据字典的必备性定义列出以下一些必备的元素，这个清单指出了保存系统应该了解的有关每一个数字对象的信息。如果数字对象的有关信息没有显示记录，那么应该能从保存系统本身或保存系统的策略、程序文档中找到这些信息。

2.3.2 对象实体语义单元

在规范中，对于对象类型是“文件”的数字对象来说，下面的语义单元是必选的。但因为它们可能不适用于所有的对象类型，所以在 PREMIS 的 xml 模式中它们可能不是必选的。

- 对象标识符类型
- 对象标识符值
- 保存级别
- 对象类型
- 组分级别
- 存储载体

此外，下面的附加语义单元在保存系统中也相当重要，应该也是必选的。

- 电文摘要算法
- 电文摘要
- 大小
- 格式名称
- 原始文件名称

2.3.3 事件实体语义单元

PREMIS 并未规定事件强制存在。一个事件可以通过对象实体的可选关系或链接事件标识符元素，链接到一个数字对象上，或者它可以通过事件的可选链接对象标识符链接到一个对象上。然而我们建议下面的语义单元是必选的：

- 事件标识符类型
- 事件标识符值
- 事件类型
- 事件日期

事件应该了解它所操作的对象。然而，PREMIS 在对象实体和事件实体中都没有指定事件集和对象集之间的必选链接。在 METS 纲要中，对象实体必须包含受缴事件的链接事件标识符，而事件实体不必包含相互的链接对象标识符。

2.3.4 代理实体语义单元

PREMIS 没有规定代理实体必须存在，因为链接代理标识符在事件实体中是可选的。然而，我们建议保存系统应该了解代理，是否事件造成了对象的改变。例如，代理应该确认用到的软件。虽然这个软件可能已经在对象实体（在创建程序中）或事件实体中加以描述，但若把它放在遵守 METS 纲要的文档实体的代理中，会促进接收典藏系统数据库的映射。如果一个组织而非迁移保存系统对一个事件负责的话，代理也应该用于组织。

2.3.5 权利实体语义单元

PREMIS 没有规定权利实体必须存在，因为链接许可声明标识符在对象实体中是可选的，PREMIS 集中考虑保存活动。然而，权利应该是必选的，只要保存系统有协议或存储系统达到一些条件，但这可能不适于资源没有版权的情况。

权利元数据对有限制访问条件的资源是必备的。如果没有记录访问权利，就假定它是无限制的。本指南不制定具体的访问权利元数据模型，只是推荐 METS 权利、PREMIS 权利、Creative Commons 许可和 XACML。

2.4 描述元数据

描述元数据在保存系统中被认为是必选的，但在本指南中不做详细介绍，可以引用 DC、MODS 等已有的描述元数据规范，以遵守 METS 纲要关于元数据交换的规定，但是应该储存并以保留所有可用元数据粒度的格式输出描述元数据。针对国家图书馆的具体实施，我们推荐使用国家图书馆已发布的描述元数据系列标准。

2.5 文件格式元数据

2.5.1 概述

文件格式元数据用来记录数字对象的特性，以便它能正确再现。在某些情况下，没有文件格式元数据，系统可能根本无法处理数字对象。参考国外已有的文件格式元数据方案，本指南制定了不同类型资源选用的不同文件格式元数据方案和扩展，这些格式包含必选元素。虽然自动获取这些元数据技术上还不可行，但对这些数据的记录能帮助我们长期管理这些数据。还需注意，该技术元数据集并不完全，具体执行还需根据具体情况进行扩展。

2.5.2 图像

MIX 是图像元数据中推荐使用的扩展模式。MIX 是一种 METS 委员会认可的模式。美国国会图书馆网络开发部和 MARC 标准办公室与 NISO 技术元数据数字图像标准委员会及其它专家合作，致力于开发一个用于管理数字图像资源的技术数据元素的 XML 模式。这个模式提供了一种格式，用于交换或和/或存储 NISO 标准数据字典中指定的数据：用于数字图像的技术元数据（版本 2.0），被称为“XML 中的 NISO 图像元数据（NISO XML）”。MIX 用万维网上的 XML 模式语言表示，由网络开发部和 MARC 标准办公室维护。

因 MIX 标准已成为美国国家标准 ANSI-NISO Z39.87(2006)，本指南推荐用 MIX 作为处理图像元数据的扩展模式。（详细 schema 见 http://www.loc.gov/standards/mix/mix.xsd）

2.5.3 音频

国会图书馆音频（源）数据字典，作为音视频资源原型项目的一部分，是推荐使用的音频元数据基本扩展模式。网页地址为 http://www.loc.gov/rr/mopic/avprot/DD_ASMD.html。下面的元数据是根据澳大利亚国家长期保存项目，对国会图书馆音频（源）数据字典的进一步扩展。

file_format

音频文件的类型，例如，WAV 文件是 Microsoft WAVE，同样 AIF 是音频文件交换格式（Audio Interchange File Format）。

file_version

所用的文件格式版本

coding_history

指明文件格式经历的历史（和设备）

mime_type

MIME 类型帮助浏览器将特定文件和合适的播放器程序或插件相联系。

compression

文件中用到的压缩类型－对没有存档复本的“非存档的”性质的文件，可能像 MPEG 压缩。

codec_version

用到的解码器版本（任选）

file_container

用于容纳其它文件格式的文件格式类型，例如 Broadcast Wave Format (BWF)就是一个可以包含 Microsoft WAV 文件的文件容器。

file_container_version

用到的文件容器版本。

frame_rate

每秒帧数。

byte_order

例如“大字节序”和“小字节序”。

timecode_type

记录在音频源项目上的时间码的类型，SMPTE（音视频同步码）丢帧，SMPTE（音视频同步码）非丢帧等。

channel_num

表示频道数，例如频道 0。这是个可重复域。

channel_num_map_loc

它与特定频道数绑定，用于表明频道的位置，例如，频道 0 的立体声文件的频道数图定位可能是“左”。

channel_map_config

频道映射的配置。在多频道工作时，这个信息非常重要。“鞋盒”和“双菱形”都是配置的例子。

delivery_type

针对流媒体文件（例如，实时流协议 RTSP）。流媒体文件不建议纳入，因为它们是非档案格式，若流媒体文件异常地被纳入，则传递协议的记录必须可用。例如，QuickTime 文件就是这种设置。“暗示流”表明这些文件只能运用 RTSP 协议访问。

encoding_softwar

运用软件来编码传递文件（只有在发生非档案性质传递文件的意外，它才是必要的）。

codec_essence

解码器用到的特殊类型或“种类”，例如 RealMedia“音乐”或“声音”解码器。

codec_essence_version

解码器实质用的版本。

2.5.4 视频

国会图书馆视频（源）数据字典，作为音视频资源原型项目的一部分，是推荐使用的视频元数据基本扩展模式。网页地址为 http://www.loc.gov/rr/mopic/avprot/DD_VSMD.html。下面的元数据是根据澳大利亚国家图书馆长期保存项目，对国会图书馆视频（源）数据字典的进一步扩展。

file_format

视频文件的类型，例如，MOV 文件是 QuickTime 文件格式，然而应该注意一种格式可能既是文件格式，又是容器格式，在视频中这种情况很常见。

file_version

所用的文件格式版本

coding_history

指明文件格式经历的历史（和设备）

mime_type

MIME 类型帮助浏览器将特定文件和合适的播放器程序或插件相联系。

compression

文件中用到的压缩类型——对没有存档复本的“非存档的”性质的文件，可能像 MPEG 那样压缩，或 MPEG 2，它是用于 DVD 的格式。当存档资料不适于压缩存储时，需注意当前阶段，视频文件非常大，因为其它的一些限制（例如大规模数据存储设施的高花费）存储非压缩视频可能不可行。视频在存档和保存方面仍是一个非开发的领域，经过一段时间以后据推测视频存档的标准会发生变化。

codec_version

用到的解码器版本（任选）

file_container

用于容纳其它文件格式的文件格式类型，例如 QuickTime(MOV)就是一个可以包含其它格式文件的文件容器。文件格式和容器格式间存在相似性。

file_container_version

用到的文件容器版本。

frame_rate

每秒帧数。

byte_order

例如“大字节序”和“小字节序”。

counting_mode

NTSC 丢帧或非丢帧。

track_num

表示轨数，例如频道 0。这是个可重复域。（类似于音频元数据域 channel_num）。

track_map_loc

它与特定频道数绑定，用于表明频道的位置。

track _map_config

频道映射的配置。在多频道工作时，这个信息非常重要。

delivery_type

针对流媒体文件（例如，实时流协议 RTSP）。流媒体文件不建议纳入，因为它们是非档案格式，若流媒体文件异常地被纳入，则传递协议的记录必须可用。例如，QuickTime 文件就是这种设置。“暗示流”表明这些文件只能运用 RTSP 协议访问。

encoding_software

运用软件来编码传递文件（只有在发生非档案性质传递文件的意外，它才是必要的）。

broadcast_standard

包括 PAL、NTSC、SECAM、DV、HDV 等。

anamorphic

一个回放设置，与 DVD 视频的控制和是否视频能在屏幕上以不同的高宽比重放有关。例如，能够在屏幕上以 4：3 和 16：9 的高宽比播放，而不致使画面压扁。可取值为“真”或“假”。

field_dominance

设为低（偶）或高（奇）。

alpha_channel

是否视频有初频道。

codec_essence

解码器用到的特殊类型或“种类”，例如 RealMedia“音乐”或“声音”解码器。

codec_essence_version

解码器实质用的版本。

2.5.5 文本（Text），HTML and XML

由纽约大学 Elmer Bobstl 图书馆的 Jerome McDonough 创建的文本技术元数据模式得到了 METS 编辑部的使用 METS 的认可，本指南推荐作为参考，不过基于国内外目前的情况，文本文档所需的附加元数据分析仍有待于进一步研究。

2.5.6 网站

保存元数据框架理论来说是不分对象类型的。但是网络资源具有特殊性。首先，网络资源是非常复杂的对象。例如，一个单独的网页会链接到多种格式的文件例如，文本、图像、音乐、多媒体、软件），这些格式都需要在保存时单独考虑。网络资源还包括数据库、数字的库文件、元数据几何和交互式网站等。另外，一些网站行为是在服务器端决定的，功能性

的其他方面依赖于浏览器软件和插件的组合。另一问题是网络资源的动态属性。web 归档机构只能保存网站和域的“快照”，随着网站的不断变化，保存的内容可能错过网站的最主要的特性。

2003 年，由 12 个国家图书馆和 Internet Archive 组成了一个国际 Internet 保存联盟（IIPC），其目标是通过国际间的合作交流，建立起 Internet 信息资源的获取和保存机构，并且使这些资源能够在未来足够长的一段时间之后仍然能够被人利用。IIPC 的一项重要工作就是支持开发和利用通用的工具、技术和标准，来构建全球的 Internet 存档。

为了实现大规模的 Web 存档，IIPC 提出了基于 OAIS 的网络资源存档系统技术体系框架。该框架覆盖了 Web Archive 工作链中的所有过程，包括采集（Ingest）、存储（Storage）、访问（Access）和索引与检索（Index & Search）等主要功能。目前技术体系框架已为大多数 Web Archive 项目所遵循和采用。

2004 年，IIPC 提出了网络资源存档的保存元数据集。在该元数据方案认为，非隔离的文档结合在一起形成内容的网络和内容的描述，因此，对于描述性信息无需太关注。而技术依赖性才是网络材料或者更一般的数字对象的关键。该方案还认为 Web archive 中选择 （时间/空间）内容是十分重要的，需要对选择的内容进行存档且将其嵌入到工具（爬行工具&测量）中。与网络服务器的每一次交互可以产生唯一的个性化的响应，需要记录这种交互或者伪交易的语境。

对于元数据的应用级别而言，元数据元素可以覆盖单个文件或若干文件。元数据应用级别可以限制在页或覆盖整个网站。

尽管 IIPC 不实施 PREMIS 数据字典，其元数据涉及对象类别，诸如“网站”、“页”和“文件”，其中 “网站”和“页”相当于 PREMIS 的表现级别，文件对应于 PREMIS 文件级别。IIPC 的另一重要成果是在 ARC 的基础上，推动并形成了 WARC 文件格式标准（ISO 28500:2009, Information and documentation -- WARC file format）。WARC 格式更好地支持采集、访问和归档组织交换的需要。除记录当前主要的内容之外，它还提供相关的次要内容，诸如分配元数据，简化副本探查，和以后的数据转换功能。WARC 支持 Heritrix，HTTrack 等抓取工具。

国家图书馆如果需要对网站进行长期保存，本指南推荐参考采用 IIPC 的网络资源存档保存元数据集。

3 保存元数据取值的自动化、规范化

3.1 概述

保存元数据实施的一个关键因素就是数据值能否由保存系统自动提供、自动进行，为此需要保存系统先定义一些命名域、受控词表，并引用一些已有的标准规范。需要说明的是，需要定义的受控词表中的建议取值并不是穷尽的枚举，国家图书馆在具体的实施中需要根据

实际情况进行补充扩展，形成符合实际需求的相关受控词表。

3.2 需要开发的唯一标识符命名域

- 对象唯一标识符命命名域
- 事件唯一标识符命名域
- 代理唯一标识符命名域
- 权利声明唯一标识符命名域

保存元数据规范中，凡涉及到标识符的语义单元，都需要相应的标识符命名域支持。

3.3 需要定义的受控词表

- 保存级别值（preservationLevelValue）受控词表
  建议取值：完全保存、比特级保存
  保存级别职责（preservationLevelRole）受控词表
  建议取值：要求、能力
- 电文摘要算法（messageDigestAlgorithm）受控词表
  建议取值：MD5、Adler-32、HAVAL、SHA-1、SHA-25、SHA-384、SHA-512、TIGER、WHIRLPOOL
- 格式名称（formatName）受控词表
  建议取值：Text/sgml、image/tiff/geotiff、Adobe PDF、unknown
- 限制类型（inhibitorType）受控词表
  建议取值：加密、密码保护
- 限制目标（inhibitorTarget）受控词表
  建议取值：对象内容、打印
- 内容位置类型（contentLocationType）受控词表
  建议取值：URI、hdl、NTFS、EXT3
- 存储载体（storageMedium）受控词表
  建议取值：磁带、光盘、硬盘
- 环境性质（environmentCharacteristic）受控词表
  建议取值：最小环境、工作环境、建议环境、不明
- 环境目的（environmentPurpose）受控词表
  建议取值：阅读、修改、转换、打印、操作
- 软件类型（swType）受控词表
  建议取值：显示软件、辅助软件、操纵系统软件、驱动程序软件
- 硬件类型（hwType）受控词表

建议取值：处理器、内存、输入/输出设备、存储器

- 签名信息编码（signatureInformationEncoding）受控词表
- 签名方法（signatureMethod）受控词表

  建议取值：DSA-SHA1、RSA-SHA1
- 关系类型（relationshipType）受控词表

  建议取值：结构、源流
- 关系子类型（relationshipSubType）受控词表

  建议取值：兄弟关系（has sibling）、零整关系（is part of）、整零关系（has part）、起源关系（is source of）、有源关系（has source）、有根关系（has root）、包含关系（include）、被包含关系（is included in）
- 事件类型（eventType）受控词表

  建议取值：收割（capture）、压缩（compression）、创建（creation）、下架（deaccession）、解压缩（decompression）、解密（decryption）、删除（deletion）、数字签名确认（digital signature validation）、传递（dissemination）、固定性检查（fixity check）、受缴（ingestion）、电文摘要计算（message digest calculation）、迁移（migration）、复制（replication）、确认（validation）、查毒（virus check）
- 事件结果（eventOutcome）受控词表

  建议取值：00（表示行为成功结束的编码）、CV-01（表示检验通过确认）
- 代理类型（agentType）受控词表

  建议取值：人、机构、软件
- 权利基本原则（rightsBasis）受控词表

  建议取值：版权、特许、法令
- 版权状态（copyrightStatus）受控词表

  建议取值：保留版权、公众领域、未知
- 行为（act）受控词表

  建议取值：复制、迁移、修改、使用、散布、删除
- 链接代理功能（linkingAgentRole）受控词表

  建议取值：联系（contact）、创建者（creator）、出版者（publisher）、权利持有者（rightsholder）、准予者（Grantor）

3.4 日期和时间格式

所有指定日期（或日期与时间）使用的语义单元都建议使用结构化形式，以辅助机器处理。数据字典因独立于实施，所以并没有指定使用某种标准。建议实施需要时采用的日期格式应符合 ISO8601[W3CDTF]规范，使用 YYYY-MM-DD 的格式。在日期不确定或有疑问的时候，建议采用约定的方式来表达一个时段，著录起讫日期时，年、月、日之间不使用连字

符，两段日期中间用连字符链接如：201006-或者 20100703-20100823。以下是可能包括日期或日期与时间的语义单元：

- 保存级别指定日期（preservationLevelDateAssigned）
- 创建日期（dateCreatedByApplication）
- 事件日期（eventDateTime）
- 版权状态确定日期（copyrightStatusDeterminationDate）
- 法令信息颁布日期（statuteInformationDeterminationDate）
- 授权开始日期（startDate）
- 授权结束（endDate）

3.5 其他语义单元取值应遵循的规范

- 版权管辖区域（copyrightJurisdiction）的取值应遵循 ISO3166 国家/地区代码
- 法令管辖区域（statuteJurisdiction）的取值应遵循 ISO3166 国家/地区代码或者其他机构名称代码

此外，国家图书馆需要建立格式注册系统，以便一组格式注册中心语义单元进行取值。建议国家图书馆在通用格式的注册选择上采用 PRONOM 作为注册中心，以实现格式信息的共享。一些私有格式通过系统设计的内部格式注册来标识。

4 实施

4.1 概述

保存元数据就是支持数字资源长期保存功能的元数据。它记录下了为实现长期保存之目的而必须记录下的技术、权利、管理等信息。这些信息之间是有规律、有逻辑的。因此，PREMIS工作组建立一个数据模型来对其进行有效组织。在这个模型里定义各种实体、为实体定义了语义单元、为实体间定义了关系。通过这个模型，可以将有益于对元数据的数据信息有序化。然后利用元数据抽取工具，保存机构可将存储库里的数字资源中所包含的保存元数据信息提取出来，再将保存元数据的信息放进这个模型里。这个模型不能只停留在概念上，它需要有一个载体，这个载体就是XML schema。利用了XML所具有的数据库功能，封装在XML文档中的保存元数据信息可以得到有效的存储、管理和利用。由此，数据模型在XML中得到了真实的实现。在保存元数据的利用过程中，又可以根据实际情况来对元数据进行完善。这样就形成了一个保存元数据与数字资源长期保存之间的一个良性循环。

4.2 数据模型的实施

PREMIS 数据模型是为了阐明数据字典中语义单元的意义和使用。它不是为了指定实施

的结构。

大多数保存系统都需要以某种方式处理概念实体、对象、代理、事件和权利，这有助于区别各对象子集的属性，如文件、比特流和表现。不过，某个保存系统的实施可能需要多或少的粒度，或定义不同实体类别。PREMIS 建议任何使用的数据模型都要清晰定义和记录，而且元数据决定和数据模型相一致。

语义单元会被分组，和某些实体间接相关。例如，environment 是对象的一个属性。逻辑上，每个文件有一个或多个相关的环境。然而，在很多情况下，环境是由文件格式决定的；也就是说，某种格式的所有文件会有相同的环境信息。不同的实施中，处理的方式可以有多种。例如：

- 保存系统 1 使用一个关系数据库系统。它有一个“文件”表格，一行对应一个文件对象，一个“环境”表格，一行对应一组唯一的环境信息。“文件”表格可以连接“环境”表格，获取每一文件的相应环境信息。
- 保存系统 2 使用一个外部维护的注册表获取环境信息。它维护一个文件格式的内部目录，以及访问外部注册表的密钥。环境信息可以通过一个 Web services 与外部注册表的接口访问，需要时可以动态获取。
- 保存系统 3 使用一个系统，把表现模拟为容器，文件模拟为容器内的对象。每一对象包括一组属性/类型值对。属性定义值的角色。属性和类型定义本身也是对象，其标识来自与其他对象标识相同的命名空间。一个文件对象可以包括一个格式属性。因为格式描述也是一个对象，它可以包括一个环境属性，依次指向一个环境描述对象。或者，一个文件对象可以直接包括一个环境属性。

4.3 元数据存储

PREMIS 实施策略工作组的调查表明保存系统存储元数据使用了若干不同的结构。最常见的是，元数据存储在关系数据库表。元数据也经常存储为 XML 数据库中的 XML 文档，或与内容数据文件一起存储的 XML 文档。实施者大部分都在使用两种或多种方法。

在数据库系统中存储元数据元素的优点是访问速度快、更新容易、查询和报告使用方便。把元数据记录作为保存系统存的数字对象和元数据描述的数字对象存储在一起也有优势：较难把元数据和内容分开，应用于内容的同一保存策略可以应用于元数据上。对实践的建议是对关键的元数据进行存储，这两种方式都使用。

复合对象需要结构元数据来描述对象的内部结构及其部分之间的关系。在 PREMIS 数据字典中，以“related”和“linking”开头的语义单元可用来表示某种简单的结构信息。但在复杂情况下，对象的表现、导航等处理经常需要用到标记丰富的结构元数据，因此，对于国家图书馆各种复杂对象的统一管理，推荐采用 METS 标准进行结构描述。这样，包含结构元数据的文件是一个文件对象，依照其本身的形式而保存。不管一个独立结构元数据的文件是否作为表现的一部分而存在，当一个存档表现输出到另一个仓储时，链接文件和表现的

元数据都应该要提供。

4.4 提供元数据值

大多数保存系统会处理大量材料，因此应该尽可能让元数据的创建和使用自动化。很多PREMIS 语义单元的值可以通过程序解析文件获取，或可以作为保存系统摄入程序的常数提供。在人工干预不可避免的情况下，实施时可以为需要代码值的语义单元配上一个允许文本解释的语义单元。

当向保存系统提交对象的个人或组织提供信息时，对保存系统的操作建议是尽可能通过程序校验这一信息。例如，如果一个文件名包括一个文件类型扩展名，保存系统就不应该假定该文件扩展名的格式自明，而应该试图在将其记录为元数据前校验文件格式。

为了便于自动处理，建议对许多 PREMIS 语义单元使用受控词表。PREMIS 假定保存会采纳或定义对其有用的受控词表。数据字典指出了在哪些地方最好的方法是要使用受控词表。国家图书馆在具体实施中可以选择所使用的词表，并说明使用的是哪个词表。是否以及如何确认使用了适当值是实施的考虑。本指南第三部分列出了操作时需要定义的受控词表。PREMIS 工作组也建立了一个机制来注册受控词表，和 PREMIS 语义单元一同使用，并将其以某种方式展示出来，让 PREMIS schema 包含进去。保存系统可以直接使用它们，或定义自己的词表，但在导出交换元数据时，应当清楚每一个受控词表的来源。如果使用和声明的是共用词表，互用性将会加强。

4.5 PREMIS 语义单元的实现

4.5.1 概述

保存系统的环境决定了它需要记录的保存元数据。实际操作中，信息记录和实现的方式更易受正在开发的系统类型的影响。使用 XML 结构作为存储和转移机制的项目和保存系统会创建基于元数据元素的系统，该系统的每个元数据被显式记录。而实现关系数据库管理系统的保存系统会在数据库结构和业务规则的设计中，隐式地记录大量保存元数据信息。关系数据库管理系统的数据模型的设计能隐式地捕捉实体间的关系信息，例如对象实体的结构或对象到环境信息的链接。关系数据库管理系统中使用的业务规则能记录所适用的广泛对象类型的信息，而该信息存放于一个元数据元素中，并被多数基于 XML 框架的保存系统中的每个对象重复。一些保存系统使用这两个方法的组合。

这两种类型的系统中，多数元数据创建方法是内部开发的，能作为一个摄入过程的一部分来执行。几乎所有的保存元数据是在数字对象生命周期的早期阶段创建的。目前只有几种现成的工具可用于实施，例如 JHOVE 和 DROID，它们主要处理技术元数据，因此许多项目用同样的工具来抽取元数据值。更多信息参见第 5 节的工具部分。

PREMIS 数据模型中的语义单元是一个实体的相关属性。一个语义单元可能是包含其它

语义单元的容器，也可能是一个具有相关值的单独的单元。语义单元的结构层次可能是一个单独的语义单元（称作单元）或一组语义单元（容器），这取决于这些单元的具体实现。通常讨论整组语义单元的应用，而不是单个的单元。适用性，必备性和重复性直接取自 PREMIS 数据字典中的规定。

在实现语义单元时，通常需解决两个问题。第一个问题是什么值要被保存到对应的语义单元里，包括要使用哪些值的决策过程，第二个问题是怎么创建这些值，从而能在系统中实施和记录。例如，对于对象标识符这个语义单元，保存系统必须决定需要的标识符类型（通常是一个策略决策），然后实现一个工具来生成这些标识符。

需要说明的是，国际上现有的保存系统中都只是部分地采用了 PREMIS 数据字典中的语义单元，因此本指南中的这部分内容也只是调查了目前保存系统中的部分 PREMIS 语义单元，分析其具体实现方法，为国家图书馆的具体实施提出建议，作为参考。

4.5.2 对象标识符（objectIdentifier）

语义单元名称：对象标识符

结构层次：容器

适用性：表现，文件，比特流

必备性：必备

重复性：可重复

实施建议：对象标识符可以在数字对象被提交到保存系统时创建，也可在系统外创建后作为其元数据和数字对象一起提交到保存系统。标识符可由系统自动生成，也可由人工分配。为了确保其唯一性和可用性，建议由系统自动生成的标识符作为主要标识符，系统外分配的标识符作为第二标识符，以便系统从数字对象链接到系统外的信息。

现有的保存系统很少提到内部标识符，因为多数保存系统能创建它们，作为一个标准特征。为了生成唯一的存档标识符，荷兰皇家图书馆（KB）实施了国家书目编码（NBN），而 Portico 采用了 John Kunze 开发的名为 NOID 的工具。国家图书馆也需要开发一个唯一标识符系统以支持本必备语义单元的取值。

4.5.3 对象类型（objectCategory）

语义单元名称：对象类型

结构层次：单元

适用性：表现，文件，比特流

必备性：必备

重复性：不可重复

实施建议：根据元数据和对象的保存需求，保存系统应可管理多种对象类型 （表现、文件、比特流）。对象类型描述了是否一些元数据适用于一个表现、文件或比特流。处理这

个语义单元有两种不同的方法，这取决于保存系统的具体实现。它或者用作一个元数据元素，或者用作一个保存系统的隐式结构特征。

在 XML 框架（例如 PREMIS：对象的 XML 框架）用于存储元数据的地方，该语义单元用作一个由受控词表填充的元数据元素，并显式地用于链接元数据到一个特定对象类型上。例如，康乃尔大学图书馆和哥廷根大学图书馆的合作项目 MathArc 在一个 XML 结构中使用“表现”和“文件”，来区分它所管理的两种对象类型。斯坦福数字保存系统（Stanford Digital Repository，SDR）打算采用包含全部三种类型的受控词表，“表现”，“文件”和“比特流”。

第二种实现方法的应用场景是该语义单元不被显示记录，而隐含在保存系统的结构中，例如在一个关系数据库中。数据模型通常关系到一个反映对象类型层次结构中的对象实体。表现实体往往链接到一个或多个文件实体，而且文件实体可以链接到零个或多个比特流实体。因此，这个层次关系通过在层次中安置实体/对象，隐式地记录了对象类型。例如新西兰国家图书馆（National Library of New Zealand，简称 NLNZ）系统就以这种方式管理表现和文件。新西兰国家图书馆的国家数字遗产档案馆（National Library of New Zealand，National Digital Heritage Archive ，简称 NLNZ NDHA）还将通过元数据管理比特流。

一个保存系统如果只管理一个级别的对象，可能不会用其它的方法来记录这个语义单元，因为它对所有对象的操作都是一样的。例如，美国国会图书馆的国家数字报纸项目（National Digital Newspaper Program，简称 NDNP）描述对象都指向文件级，并不明确记录对象类型。

4.5.4 保存级别（preservationLevel）

语义单元名称：保存级别

结构层次：容器

适用性：表现，文件

必备性：可选

重复性：可重复

实施建议：保存系统可以描述适用于它们对象的保存级别。但不同系统对保存级别的表达却不同。一些保存系统认为保存系统的目的就是为某种类型的对象（例如，“是数字原件”（isDigitalOriginal））提供保存服务。其它保存系统则使用一些反映当前存储能力的术语来保存对象的格式（例如，“比特级”）。根据保存系统的目的，可以选择是把系统内的所有资料（material）记作一个保存级别，或者为对象的一或多个属性提供几种有限选择。

有两种类型的保存系统可能选择单个的保存级别。第一种类型由于它们所获取的对象类型，具有受限的保存前景。例如美国国家冰雪数据中心（National Snow and Ice Data Center，USA ，简称 NSIDC），它主要收集科学数据集，并把这些数据描述为只能以比特或字节级保存。数据本身可能并不具有表示特征。因此，只需决定这些数据（以及它的附属文档和元

数据）是否应保存，而不用决定它们应该以什么级别保存。然而，NSIDC 实际上不但记录保存级别，而且评估其是否有其它的适用条件，以备将来开发时进行保存级别决策。

第二种类型目前保证只适用于单个保存级别，它同等看待所有的内容，并且无论什么对象都保证应用同一级别。例如英国国家档案馆（The National Archives， UK，简称 TNA），它旨在为所有的对象应用相同级别的保存承诺（preservation commitment）。另一个例子是荷兰皇家图书馆（Koninklijke Bibliotheek，简称 KB），它认为所有的资料（material）是最重要的，并保证保留所有对象的“观感”。在上面这两个例子中，保存系统都只选择一种级别的保存承诺，它们可以只在策略级别记录决策，而不需要在每个对象的元数据中都记录决策。

保存级别最常见的应用是为保存系统从一个替换定义的保存承诺（preservation commitment）级别中选择一个值。下表包含了这些级别的一些取值的例子，国家图书馆可以在定义自己的保存级别取值受控词表时作为参考。

保存级别术语（terms）示例

保存系统	保存承诺		
	高级	中级	低级
新西兰国家图书馆（NLNZ）	是保存原本	是数字原件	是访问复本
英国 SHERPA 人文数据服务数字保存项目（SHERPA DP）	00（完全）	01（仅内容）	02（比特级）
Portico	完全支持	适当措施（Reasonable effort）	字节保存
Florida Digital Archive （FDA）	完全（完全保存）	比特（仅比特级）	无（不存档）
Deep Blue Michigan	1 级（最高）	2 级（限制）	3 级（目前状况）

理想情况下，保存级别的选择可以基于一系列标准自动执行，这些标准是明确定义的，并且根据保存系统的业务规则进行编码。为了自动化，最安全的选择可能是把一个保存系统的最高级保存承诺作为默认值，除非另有说明。下表列出了几个保存系统决定其保存级别的标准，国家图书馆需根据自己的保存策略确定自己决定保存级别的标准。

决定保存级别的标准示例

保存系统	决定保存级别的标准
NLNZ	存储细节，对象生命周期
SHERPA DP	适于保存的文件格式，保存权利
Portico	格式和格式有效性
FDA	存储者帐户协议

Deep Blue Michigan	文件格式预期寿命

参考目前的一些保存系统，国图可以定义三个保存级别：完全保存、比特级和不保存，对过程数据不进行保存。并对不同类型的资源定义不同的保存级别，具体可参考如下：

1. 文本类型

文本格式	保存级别
PDF	完全保存
SGML	完全保存
XML	完全保存
HTML	完全保存
TXT	完全保存
DOC	完全保存
PPT	完全保存
XLS	完全保存
国图自有格式	完全保存
其他格式	比特级或者不保存

2. 图片类型

图片格式	保存级别
JPG	完全保存
TIFF	比特级
PNG	比特级
BMP	比特级
GIF	比特级
PSD	完全保存或比特级
其他格式	不保存

3 音频类型

音频格式	保存级别
WAV	完全保存
MP3	完全保存
WMA	比特级或不保存

4. 视频类型

视频格式	保存级别
AVI	完全保存
MPEG	比特级

WMV	不保存

4.5.5 重要属性（significantProperties）

语义单元名称：重要属性

结构层次：容器

适用性：表现，文件，比特流

必备性：可选

重复性：可重复

实施建议：目前所有的长期保存项目中还没有一个完全为该语义单元开发的词表，而且它们以不同的级别使用重要属性，从文件级到语义实体资源。因此需要更多的数字保存经验，来确定最好的表示重要属性和重要属性修改的方法。

美国国会图书馆的国家数字报纸项目（National Digital Newspaper Program，简称 NDNP）使用重要性质来记录任何文件可违反的规则。也就是说，NDNP 已经为所有文件格式制定出详细的纲要，而且在某些情况下，这些纲要可以违背。NLNZ NDHA 在保存策略中记录与表现级别相关的重要属性，而不太可能记录在各对象中。KB 不打算以对象级别记录重要属性。它们把重要属性与一类对象或“资源”相关联。英国 SHERPA 人文数据服务数字保存项目（SHERPA Digital Preservation Project at the Arts and Humanities Data Service， UK，简称 SHERPA）专门构造重要属性以适用电子预印本。一个电子预印本的重要属性能识别一些特别的属性，这些属性必须通过以后的保存动作（例如迁移）加以维护，或可对保存动作产生影响。一个电子预印本的重要属性的最常见的例子包括语义内容（文本和图片），以及文档布局。英国国家档案馆（The National Archives， UK，简称 TNA）描述两个概念上不同类型的属性，它们对数字对象非常重要。它们讨论了将语义实体的不变属性和表现的技术属性进行混合。不变属性是那些对真实性至关重要的记录属性，它必须能保存较长的时间，并不受不同表现形式限制。TNA 以语义实体级别联合这些属性，因为它们关系到概念记录。TNA 打算通过分析组成文件，测度任何给定表现的不变属性。这些测度会允许迁移结果的验证，并提供这些属性的容许公差的定义。技术属性随着一个表现的每个表现形式而改变。Stanford 使用适用于所有格式的技术元数据的重要属性，但不包括某些格式技术元数据框架。NDNP，SHERPA，TNA 和 Stanford 都为重要属性提供了一些专有格式对象特征的技术属性。或许这造成了一个误解，PREMIS 到底打算利用重要属性解决什么问题，但这肯定会指明在专有格式保存元数据方面的未来工作需求。

国家图书馆在具体实施长期保存的过程中，可以对此语义单元进行扩展，增加“文献类型”，用“文献类型”语义单元首先确定长期保存数字对象的类型，再分别对不同类型的文献定义各自的重要属性类型和值。“文献类型”语义单元的取值可以是文文本、图片、音频、视频等。具体扩展参考如下：

1.4 重要属性（significantProperties）

1.4.1 重要属性类型（significantPropertiesType）

1.4.2 重要属性值（significantPropertiesValue）

1.4.3 重要属性扩展（significantPropertiesExtension）

1.4.4 文献类型（documentType）

1.4.4.1 文本（text）

1.4.4.1.1 格式（textFormat）

1.4.4.1.2 字符集（characterSet）

1.4.4.1.3 语言（language）

1.4.4.1.4 内容（content）

1.4.4.1.5 页数（pages）

1.4.4.1.6 页宽（pagewidth）

1.4.4.2 图片（image）

1.4.4.2.1 格式（imageFormat）

1.4.4.2.2 色彩（color）

1.4.4.2.3 像素（pixel）

1.4.4.2.4 压缩率（compressionRate）

1.4.4.3 音频（audio）

1.4.4.3.1 格式（audioFormat）

1.4.4.3.2 音色（toneColor）

1.4.4.3.3 信号模式（signalMode）

1.4.4.3.4 比特率（bitRate）

1.4.4.3.5 信噪比（signalNoiseRatio）

1.4.4.3.6 音频编码方式（audioEncodedMode）

1.4.4.4 视频（vedio）

1.4.4.4.1 格式（vedioFormat）

1.4.4.4.2 帧数（frames）

1.4.4.4.3 扫描模式（scanMode）

1.4.4.4.4 分辨率（resolutionRatio）

1.4.4.4.5 像素格式（pixelFormat）

1.4.4.4.6 色彩空间（colorSpace）

1.4.4.4.7 视频品质（vedioQuality）

1.4.4.4.8 位元传输率（bitTransferRate）

1.4.4.4.9 视频编码方式（vedioEncodedMode）

4.5.6 对象特征（objectCharacteristics）

语义单元名称：对象特征

结构层次：容器

适用性：文件，比特流

必备性：必备

重复性：可重复

实施建议：文件或比特流存在一些重要技术属性适于任何格式的对象。对象特征中的所有元素是适用于一个组分级别的一个对象的信息集合，对于两个或多个编码程序协作（比如压缩和加密）产生的对象，其对象特征可重复，每重复一次将增加一个组分级别。一个加密对象，其对象特征必须包含一个必备元素。文件内嵌的比特流的对象特征不同于文件的对象特征，如这些特征有助于对象保存，则需记录。当一个单独文件与一个表现形式等价时，可采用对象特征并与表现形式相关联。在这种情况下，组成表现的文件和其它相关的文件可能用关系子类型表示。

对象特征语义单元包括组分级别、固定性、大小、格式、创建程序、限制信息几个语义组分，下面分别分析其在几个系统中的实现情况。

4.5.6.1 组分级别（compositionLevel）

语义单元名称：组分级别

结构层次：单元

适用性：文件，比特流

必备性：必备

重复性：不可重复

实施建议：组分级别一般由系统自动赋予，对于保存系统创建的对象，其组分级别必须由创建程序记录并形成元数据；对于呈缴来的对象，系统须从对象中识别出其组分级别或从外部元数据中获取。一个文件或比特流可依赖于多个编/解码程序。比如，文件A被压缩后形成文件B，文件B被加密后形成文件C。如果想恢复得到文件A，首先需要将文件C解密形成文件B，然后将文件B解压缩，从而得到文件A。

组分级别排列从低到高，第一级为“0”，“0”级是基础级别，表示该对象是最基本的对象，不能再进行任何解码操作，组分级别“1”和更高的组分级别，说明该对象需要一个或多个解码程序来恢复成基本对象。如果系统仅有一个组分级别，那么“0”作为默认值。

当多个文件（作为文件流）被封装到一个文件包时（比如一个ZIP文件），每一个文件对象都不是一个文件包的组分级别，他们应该被认做是分开的不同的文件，每一个文件都有其组分级别。比如，对于两个被加密的文件压缩成的一个ZIP文件，系统需要分开描述三个不同的文件，每一个文件附带其元数据。那两个加密的文件的存储位置（storage location）需指向ZIP文件，但ZIP文件只能有一个组分级别“0”，它的格式是“zip”。

多数保存系统似乎想记录对象的组分级别，无论它们要获取捆绑的对象还是加密对象。保存系统可能不愿用压缩或加密来存储对象，因此，它们把此项记为一个业务规则而非对每个对象都记录此项，以满足这个语义单元的必备要求（必备意味着档案必须“知道这个信息”）。有的保存系统的策略是不存储压缩或加密的对象但要确认它们的XML框架，例如MathArc，这样的保存系统可能把这个必备语义单元记为缺省值“0”。

4.5.6.2 固定性（fixity）

语义单元名称：固定性

结构层次：容器

适用性：文件，比特流

必备性：可选

重复性：可重复

实施建议：固定性用来校验一个数字对象在系统没记录或未授权的情况下是否被改变的信息，由系统自动计算并记录。多数应用程序使用至少一个校验和算法来计算一个电文摘要，其中最流行的是 MD5 和 SHA－1 算法。现有的长期保存系统中，NLNZ NDHA 和佛罗里达数字存档（Florida Digital Archive，FDA）都既使用 MD5，也用 SHA－1 校验和，只有 Portico 这个系统使用 SHA－512。国家图书馆可参考已有系统，采用一种或两种校验方式。TNA 还创建了另一个元素“固定性方法”来补充电文摘要算法（固定性类型）。因此，固定性类型描述了所用算法的类型（例如，“MD5 摘要算法”），而固定性方法描述了产生信息摘要所用的工具（例如，“MD5 Summer 1.1.0.22”）。其固定性类型使用一个受控词表。

保存系统可接受附带电文摘要的文件，通过比对系统就可知道接受的文件是否就是被提交的文件。保存系统也可接受不附带电文摘要的文件，但须在接受文件时执行校验算法生成初始的电文摘要。系统记录电文摘要的创建者有利于保存管理。电文摘要语义单元在所调查的保存系统中普遍使用，并遵守 PREMIS 数据字典中的定义，而只有一半的保存系统记录了电文摘要创建者（messageDigestOriginator）。为了实现自动化，建议在使用中，或能自动添加，或在受控词表选取。

另外需要注意的是，目前所进行的长期保存项目都只在文件级使用校验和，没有一个是在比特级使用的。校验和通常在摄取工作流阶段计算，或一个校验和在保存系统接收前就被创建，而在摄取过程中进行检查。例如，FDA 中任何在摄取时提供的文件校验和都要进行验证，如果不匹配就拒绝该对象。使用校验和算法的保存系统可以选择是否在元数据元素中记录电文摘要算法，并规定为一项业务规则。

4.5.6.3 大小（size）

语义单元名称：大小

结构层次：单元

适用性：文件，比特流

必备性：可选

重复性：不可重复

实施建议：文件或比特流的比特大小可用来确保对象被正确获取，也可用来告知一个系统应用是否有足够的空间来移动或处理文件。尽管是可选的语义单元，现有的保存系统都记录了文件的大小，以字节为单元。KB 还记录了“表现”的整体大小，虽然 PREMIS 数据字典中通常认为该级别的大小不适用。捕捉文件大小是系统的常见功能，一般被加入到摄取工作流中。例如，苏格兰国家档案馆数字数据存档项目（National Archives of Scotland， Digital Data Archive，简称 NAS DDA）提出使用 Visual Basic 功能来产生该元素。

4.5.6.4 格式（format）

语义单元名称：格式

结构层次：容器

适用性：文件，比特流

必备性：必备

重复性：可重复

实施建议：PREMIS 需要记录格式标记（formatDesignation）或格式注册中心（formatRegistry）语义单元。建议除记录格式标记，例如 MIME 格式外，还应该记录格式的版本信息。一些保存系统开发了复杂的格式识别规则，有些系统只采用最简单且可能不精确的方法。澳大利亚可持续保存联盟（Australian Partnership for Sustainable Repositories，简称 APSR）推荐同时使用格式标记和格式注册中心，以防注册失败或在需要时不可用，认为本地记录该值可为报告或管理功能提供有用信息。Portico 在识别过程中使用 MIME 类型，也记录了更多的格式信息。TNA 使用 DROID 来识别文件格式和版本，它结合了内部和外部签名，而且分配一个 PRONOM 唯一标识符（PUID）用于存储，这个标识符等价于格式注册中心表。PUID 作为指针指向 PRONOM 中具体格式和环境信息。NLNZ NDHA 使用 NLNZ 元数据抽取工具来识别资源中最常见文件的格式和版本。文件的主要格式会附带文件级别相关元数据。如果抽取出附加的比特流元数据，就添加一条新元数据值来记录比特流信息。

系统需在接受文件或比特流时确定其格式，这可直接从提交者提供的元数据来确定，也可从其文件扩展名来识别。建议系统尽可能地采用中立的方法，分析对象后确定其格式。如在接收对象时无法确定其格式，就需先将其格式记录为“未知（unknown）”，然后系统需尽量识别其格式，包括通过人工干预的方法来确定。

4.5.6.4.1 格式注册中心（formatRegistry）

语义单元名称：格式注册中心

结构层次：容器

适用性：文件，比特流

必备性：可选

重复性：不可重复

实施建议：要实现数字资源的长期保存，需要开发和维护一批格式注册中心，还应建立基于网络的全球数字格式注册中心（Global Digital Format Registry），实现格式信息的全球共享。

现有的保存系统有使用本地格式注册系统的，也有使用外部格式注册中心的。KB使用自己开发的内部格式注册中心，但它并非链接到对象元数据，而是直接嵌入保存管理系统中。详述见“环境”语义单元。APSR更喜欢使用普遍适用和综合的注册中心，并允许提供多个注册中心的链接。Portico基于与全球数字格式注册中心（Global Digital Format Registry，GDFR）建立链接的目的进行开发。TNA使用DROID来提供PUID作为格式注册中心值，链接到PRONOM注册中心，不再需记录格式注册中心名称或角色。NAS DDA打算使用PRONOM作为它的注册中心，而且格式注册中心域由DROID填充。格式注册目的是希望对象格式可以在多个地方应用共享，并为人所知，但也可能某种格式的数据是某些特定软件产生出来的，例如国家图书馆的地方志格式。对这种格式的详细信息以及相关支持软件的记录会有助于特定格式的理解和保存。因此建议国家图书馆在通用格式的注册选择上采用PRONOM作为注册中心，以实现格式信息的共享。一些私有格式要通过系统设计的内部格式注册来标识。

4.4.6.5 创建程序（creatingApplication）

语义单元名称：创建程序

结构层次：容器

适用性：文件，比特流

必备性：可选

重复性：可重复

实施建议：创建程序的版本和创建日期等信息，对系统解决问题是有用的，比如某些版本的软件会带来格式转变错误或产生衍生数据。本组语义单元既适用于系统外创建的对象，也适用于系统内创建的对象（比如通过迁移）。如果对象是由系统创建的，创建程序信息需由系统直接赋予。如果对象是在系统外创建的，那么创建程序信息应该由提交者提供。系统也可以从对象文件中萃取创建程序信息，因为创建程序的名称经常是内嵌在文件中的。创建程序是可重复的，如果多个程序处理了对象，比如一个Microsoft Word的doc文件被Adobe Acrobat转化成PDF文件，需同时记录Word和Acrobat的详细信息。如果系统同时保存这两个对象，每一个对象都应该作为一个对象实体来描述，并通过关联信息中的关系类型（relationshipType）的“derivation”来实现关联。作为可重复的语义单元，可用来记录对象被提交前的创建程序，也可以用来记录收缴过程中使用的创建程序。比如，一个HTML文件在提交到系统前是由Dreamweaver创建的，由网络蜘蛛Heritrix收割并形成一个网页快照，而这一过程是收缴过程的一部分。这里仅提供创建程序的最基本的信息，可仿照环境语义单元来设计。每个保存系统可不必本地记录这些信息，最好是建立一个类似格式或环境的

注册中心。

4.5.6.6 限制信息（inhibitors）

语义单元名称：限制信息

结构层次：容器

适用性：文件，比特流

必备性：可选

重复性：可重复

实施建议：限制信息由系统在接收对象时获取，并不是由系统自动提取。一般来讲，不能通过分解一个文件来断定其是否被加密，因为文件可能是 ASCII 文本。因此，限制信息应由提交者作为对象元数据的语义单元，在提交时和对象一起提供。

许多保存系统不记录限制信息。例如，KB 规定如果对象包含限制信息，那么不允许对象提交，因此记录为空。TNA 把限制信息记录在对象的重要属性里，而不是作为一个单独的元数据语义单元。FDA 在摄取过程中记录所发现的限制信息，但作为格式确认事件的值来存储该信息，而不是作为对象的属性。FDA 不记录限制目标或限制口令。AHDS SHERPA 项目期望接收少量具有限制信息的电子预印本，该限制信息是控制访问的主要手段。他们更喜欢保存的主文件是没有限制信息的版本，但保留该限制信息，以作为迁移或转化对象时的一个重要属性。他们提议的限制类型的受控词表列出了加密或密码保护的特定类型，并反映了 PREMIS 数据字典中的列表：

－DES 加密

－PGP 加密

－Blowfish 加密

－128－bit RC4 密码保护

－证书保护

4.5.7 原始文件名称（originalName）

语义单元名称：原始文件名称

结构层次：单元

适用性：表现，文件

必备性：可选

重复性：不可重复

实施建议：原始文件名称一般由提交者提供或由收割程序确定，但文件路径（filepath）的确定由系统来确定。原始文件名称是 SIP 中的文件名称，文件可在不同的语境（contexts）中拥有其他的名称。当两个保存系统交换内容时，接收系统应该知道并记录该表现形式在原始系统中的名字。如果交换的是表现形式，那可能需记录一个目录名。

多数保存系统在摄取过程中获得原始文件名称。这个功能可以直接利用标准文件管理功能自动化。

4.5.8 存储（storage）

语义单元名称：存储

结构层次：容器

适用性：文件，比特流

必备性：必备

重复性：可重复

实施建议：保存系统不能对所管理的内容失去控制，保存系统需要通过程序来分配内容位置（contentLocation）。如果保存系统使用对象标识符作为提取数据的句柄（handle），内容位置（contentLocation）是潜在的，系统无需记录。系统要知道内容位置的值，首先需知道对象保存使用的位置编码方式（location scheme）。它可以是完全可靠的路径和文件名，也可是解析系统（resolution system，比如handle）或存储管理系统中的信息。对比特流或文件流来说，它可能是参考点和比特流的偏移量。另外，保存系统应该决定记录的粒度大小，还需知道对象存储的载体，以便于决策何时如何进行载体更新和载体迁移。虽然某些情况下，存储载体可由存储管理系统（storage management systems）管理，但保存系统强调控制，而且还需管理技术过时（technological obsolescence）。

虽然所有保存系统目前都知道怎样定位它们的对象，但很少会在元数据中显式地记录这些值。NLNZ NDHA为文件分配一个定位值，该过程由存档系统管理。当比特流信息被提取出来时，也希望元数据提取器或格式识别工具能记录文件偏移或比特流长度，以定位比特流。内容位置类型和存储载体被认为是隐含在系统中的，而不是显式地记录在对象元数据中。FDA记录文件的内容位置值和比特流的内容位置类型和值。这些值由摄取过程创建。存储载体由系统所知，并被名为“TSM”（Tivoli Storage Manager）的当前系统参考，据此可推断出磁带单元。KB能根据一个对象的功能（例如，存储或访问），推断这个对象是否光存储或磁带存储。

4.5.9 环境（environment）

语义单元名称：环境

结构层次：容器

适用性：表现，文件，比特流

必备性：可选

重复性：可重复

实施建议：环境是用户和数字内容交互的手段和方法。数字内容离开了其存在的环境将失去作用。这个语义单元的语义组分都是可选的。如保存系统仅采取比特级的保存策略，

则可省略环境信息。建议像格式注册中心那样建立一个环境信息的注册中心。如果每一个对象所需的环境和由其构成的表现所需的环境相同，则系统不必保存每一个对象的环境信息，可通过建立继承机制（mechanisms for inheritance）实现。

环境信息很少记录在PREMIS数据字典规定的扁平结构中。项目KB和TNA在开发一个用于处理该信息的系统。这是由环境组件的复杂性和持续变化发展起来的，以支持数字对象的使用。这两个系统都将对象的格式信息和技术需求（包括软件和硬件）联系起来。

KB将其对象的格式信息以表现级联系到存储在“保存管理器（Preservation Manager）”系统中的环境信息上。保存管理器使用一个由“保存层模型”（Preservation Layer Models，PLM）和“观察路径”（View Paths）组成的结构，来注册存储在其保存系统中的文件格式信息。PLM描述了文件格式怎样关联到运行在系统不同概念层上的软件和硬件。这些层表示类似于PREMIS软硬件和相关性语义单元的概念。数据格式是高层，下面各分层是每个软件应用程序组件，操作系统和所需参考平台。每一层的描述包括一些属性，例如“名字”，“版本”和“补丁”。一个“观察路径”是与一种文件格式相关的保存层模型实例。最好每个文件格式有多个“观察路径”。这意味着可以有多种方式来访问一个格式，从而增加生存寿命。观察路径的一个特定的好处是可以引起技术的变化，这会反映在一个“观察路径”的创建和折旧（deprecation）上，而不必更新对象元数据。

类似地，TNA正在开发PRONOM注册中心，以提供与格式相关的环境信息。访问对象所需的技术环境在PREMIS数据字典环境信息中给予了类似地描述，但并不直接存储在对象元数据中。一个对象的格式采用一个PUID（PRONOM唯一标识符）来进行描述，它指向PRONOM中的详细描述。在PRONOM内，格式信息会提供使用对象所需的软件信息。

4.5.10 签名信息（signatureInformation）

语义单元名称：签名信息

结构层次：容器

适用性：文件，比特流

必备性：可选

重复性：可重复

实施建议：保存系统可在收缴对象时为其附加数字签名，也需要存储并确认数字签名。国外现有的保存系统中只有一个声明使用数字签名，即国会图书馆的 NDNP，而且在 PREMIS 数据字典完成并采用 W3C 的《XML 签名语法和处理》（XML-Signature Syntax and Processing， XML 签名）标准来编码数字签名，并往其 METS 记录中添加元数据之前，该系统就实现了这些签名。这个标准比 PREMIS 签名语义单元更详细。国家图书馆在处理签名信息时，可参考 NDNP 系统。

4.5.11 关系信息（relationship）

语义单元名称：关系信息

结构层次：容器

适用性：表现，文件，比特流

必备性：可选

重复性：可重复

实施建议：保存系统需知道如何将对象的各组成部分（结构关系）进行数字溯源（derivation relationships）后，恢复成复杂的数字对象。记录数字对象的关系是实现这一目标的基本要求。大多数保存系统需记录所有数字对象的关联信息。在复杂场景中，PREMIS未必能表达足够丰富的结构关系，以作为结构元数据的唯一来源。多数表现结构信息的格式都可用来代替在此定义的语义单元。这些信息必须可获知。文件层次的结构关系在重构一个表现时是必要的，用以实现表现的应用。表现层次的结构关系也是表现显示或应用所需的。比特流层次的结构关系可将一个文件内的多个比特流关联起来。文件和表现层次的关系对于记录数字源流是非常重要的。关于关系信息的具体取值建议见本指南第三部分。

关系信息总是保存系统中一个复杂的问题，这反映在所调查的保存系统中它们的多种实现方式上。PREMIS 描述了两种类型的关系，组件间的“结构”关系以及表示履历信息的“源流”（derivation）关系。采用 PREMIS 建议的关系子类型的项目只使用了数据字典所列值的一个子集。所选择的子集在每个项目中各有不同。

现有的保存系统中有四个目前不记录任何特定关系信息，虽然它们所用的存储结构可能把具有结构关系的对象组合到一个信息包中。MathArc 也使用 METS 中的 structMap 段来表示一些（不是全部）关系。关系信息语义单元只被用来存储履历信息的源流（derivation）关系。关系类型总是“源流”，而且子类型总是“有前任”（has predecessor），因为只有在创建/迁移事件发生时，才提供后向链接。源流关系必须带一个可链接的事件，事件标识符只需要存储在包含事件信息的 METS 文件中。FDA 关系是从表现或比特流单向指向文件。表现通过一个“整零”（has part）关系关联到文件，而比特流通过“零整关系”（is part of）关联到文件。因此，文件间的兄弟关系可由此推断出来。在 TNA 中的关系数据库系统中，结构关系隐含在数据模型设计中，而且不需要子类型。关联对象标识符值只显式地进行记录。事件序列也是隐含的，依靠日期和数据库结构来维护这些关系。

4.5.12 链接事件标识符（linkingEventIdentifier）

语义单元名称：链接事件标识符

结构层次：容器

适用性：表现，文件，比特流

必备性：可选

重复性：可重复

实施建议：它用来链接那些不派生对象关系的事件，比如，格式确认和病毒扫描等。多数保存系统的事件标识符类型取自于内部的编号系统，它可以是潜在的，仅在对外输出数据时提供。

通常事件标识符局限于本系统，因此它可作为一个已知的局部标识符类型，而不必对每个对象都显式记录。多个项目间会混合使用显式和隐式记录事件标识符值，它们的使用只在文件级或者在表现和文件间，这取决于事件怎样联系到对象上。现有项目没有使用等价于这些语义单元的元素，显式地把事件和比特流关联起来。

4.5.13 链接知识实体标识符（linkingIntellectualEntityIdentifierValue）

语义单元名称：链接知识实体标识符

结构层次：容器

适用性：表现，文件，比特流

必备性：可选

重复性：可重复

实施建议：用来链接到与对象相关的知识实体，链接指向知识实体或其可被参考的代用品的描述元数据，可以链接到元数据所描述的比数字对象较高概念层的一个对象的标识符，比如，一个资源或上级对象。

现有的系统中只有 FDA 明确使用该语义单元。保存系统从一个受控词表中选择标识符类型，而标识符的值根据存储者（depositor）所提供的元数据填充。NLNZ NDHA 在数据库结构中隐含该值。TNA 期望来自知识实体的链接能显式地表示，而非来自对象。一些其它的保存系统不打算使用该语义单元。

4.5.14 链接权利声明标识符（linkingRightsStatementIdentifier）

语义单元名称：链接许可声明标识符

结构层次：容器

适用性：表现，文件，比特流

必备性：可选

重复性：可重复

实施建议：通常权利协议适用于一组对象，而不明确记录在单个对象中。详细信息见权利实体部分。

现有的保存系统中目前没有一个实现了这个语义单元，但有两个系统声称可能用到。NAS DDA 和 APSR 提出要使用链接标识符。NAS DDA 会链接表现到权利记录上，并期望权利能平等地适用于一个表现内的所有对象。

4.5.15 事件标识符（eventIdentifier）

语义单元名称：事件标识符

结构层次：容器

必备性：必备

重复性：不可重复

实施建议：保存系统中的每一个事件必须具备一个唯一标识符，并通过它实现与对象、代理、和其他事件的关联。事件标识符可由系统自动生成，目前尚不存在事件标识符的全球框架或标准。该标识符是不可重复的。

MathArc，Stanford 和 FDA 显式地记录一个事件标识符，APSR 推荐这样使用。在 MathArc 和 FDA 中事件标识符由标识符类型和标识符值组成。FDA 记录一个局部常量作为标识符值（“FDA”），以及其取值。MathArc 中，标识符类型和值都根据项目参与者的保存系统使用。标识符只需要在每一个 METS 流内部是唯一的。这个 MathArc 事件标识符被用来把一个关系链接到一个事件上。Portico 和 TNA 使用一个与 PREMIS 模型不同的结构，而且它们不需要事件实体有一个显式的标识符。

4.5.16 事件类型（eventType）

语义单元名称：事件类型

结构层次：单元

必备性：必备

重复性：不可重复

实施建议：区分事件类型有助于系统处理事件信息，特别有助于生成系统报告。PREMIS 数据字典提供了一些事件类型的建议取值，下表列出了各个系统中事件类型的受控词表，国家图书馆可以据此作为参考，定义自己的事件类型受控词表。

各保存系统中事件类型的受控词表

PREMIS	Partico	MathArc	SHERPA DP	FDA
捕捉（capture）			捕捉（capture）	
压缩（compression）				
下架（deaccession）			下架（deaccession）	
解压缩（decompression）				
解密（decryption）				
删除（deletion）		删除（deletion）	删除（deletion）	DEL（删除的文

				件）
数字签名确认（digital signature validation）				
传递（dissemination）				D（disseminated，传递的）
固定性检查（fixity check）	事件校验和确认（EvenChecksum-Verified）		固定性检查（fixity check）	VC（verified checksum，确认的校验和）
摄取（ingestion）			摄取（ingerstion）	I（ingested，摄取的）
电文摘要计算（message digest calculation）	事件校验和计算（EventChecksum-Computed）		电文摘要计算（message digest calculation）	
迁移（migration）		迁移（migration）	迁移（migration）	M（migrated，迁移到）
标准化（normalization）	事件转换文件（Event-TransformedFile）			N（normalized to，标准化到）
复制（replication）	事件数据复制（EventDatCopied）			RM（refreshed media，更新载体）
确认（validation）	事件格式证实（EventFormat-Verified） 事件格式证伪（EventFormat-VerFailed）		确认（validation）	
查毒（virus check）	事件病毒扫描（EventVirus-Scanned）		查毒（virus check）	CV（checked for virus，检查病毒）
	事件格式识别			

	（EventFormat-Identified）			
	事件 Tmd 提取（EventTmdExtracted）			
	事件状态失活（EventStatusInactive）			
	事件保存级别改变（EventPreservation LevelChanged）			CPD/CPU（changed preservation level downward /upward，向上/下改变保存级别）
	事件文件增加（EventFileAdded）			
	事件文件创建（EventFileCreated）			
	事件格式改变（EventFormatChanged）			
		替换（Replacement）		
		更新资源元数据（UpdateAsset-Metadata）		
		不一致性发现（Inconsistency-Discovered）		
			重发请求（Resub_request	

			）	
				DLK（down-loaded link，下载链接）
				L （localized to，局限于）
				WA（withdrawn by archive，从档案中提取）
				WO（withdrawn by request of owner，根据拥有者的请求提取）
				Unknown（不知道）

4.5.17 代理标识符（agentIdentifier）

语义单元名称：代理标识符

结构层次：容器

必备性：必备

重复性：可重复

实施建议：代理实体集成了数字对象的生命周期中，和权利管理和保存事件相关联的代理（人、机构、软件）的属性和特征的信息。所有的代理信息用来准确确定一个代理。唯一的必备性语义单元是代理标识符。

许多保存系统使用某些形式的代理实体。然而，它们对代理的具体实现各不相同。这通常是因为同一个组织中的其它系统也包含代理实体，例如人和组织。这正是 PREMIS 所期望的，也是 PREMIS 数据字典中的代理实体只有很少的详细信息的原因。

TNA，NLNZ NDHA 和 FDA 都在系统的其他地方记录代理信息，但最好把它映射到 PREMIS 代理实体。Portico，MathArc，APSR 和 NAS DDA 使用专用的代理实体，以供保存元数据和系统的其它部分使用。APSR 和 NAS DDA 描述了包括软件在内的可能代理类型列表。正如 PREMIS 数据字典中所列出的，它们建议使用代理来记录人，组织或软件。

根据实际的应用情况，代理实体可以扩展一个语义单元：代理职能（Role）.通过此语义单元可以更详细地说明每个代理类型的不同职能，比如代理类型为人，其职能可以取值为用户、出版者、供应者、元数据加工人等。建议为此语义单元建立受控词表进行取值。

4.5.18 权利声明（rightsStatement）

语义单元名称：权利声明

结构层次：容器

必备性：可选

重复性：可重复

实施建议：权利是版权法或其他的知识产权法律规定的代理所享有的权利。一个保存系统可能需要记录一些权利信息，这包括适用于外部代理和外部数字对象的权利声明和许可声明。一个保存系统需要知道的最小范围的核心权利信息是，保存系统对其所保存的数字对象可采取的被授权的保存行为。

各保存系统间对权利的处理也很不相同，而且可能区别于 PREMIS 数据字典中的权利实体。NAS DDA 提出使用所有对应于权利语义单元的元数据元素，该实体会链接到表现级。目前 Portico 为权利元数据提供一个位置标识符，而且它只是把元数据链接到与内容相关的存储者（depositor）合约上。该合约或协议也存储在系统内。NLNZ NDHA 会生成一个许可声明，该声明由存储者（depositor）或基于保存系统业务规则（例如，图书馆策略）手动填表输入的信息产生。通常这是一个保存终端的自动过程。TNA 记录知识属性实体，并描述访问条件，该访问条件与知识对象级的记录、结束、干扰内容等相关。这个信息只确认版权所有者和任何限制信息。MathArc 仅使用基于 OAI 资源的权利信息。在使用权利上的另一个不同是链接权利到一类对象，而不是单个对象，例如 SDR。所以，权利实体不必链接到单个对象上，然而所有的 PREMIS 语义单元被用于记录除了授权代理（grantingAgent）外的权利信息。SDR 使用一个模板来填充权利元数据，该模板提供一些值和受控列表。

5 元数据自动抽取

5.1 概述

长期保存元数据可能来自于已存的记录、策略和文档，被存储器支持，且被记录为保存过程的一部分，也可能是从资源中提取的。考虑到需求或可用的元数据数量，元数据获取过程最好自动化，尤其是从资源中提取的元数据。元数据自动抽取的一般过程是：首先，对数据来源进行必要的预处理，剔除在格式、内容等方面存在问题或严重缺失的文档；其次，经过元数据抽取模块的处理，生成符合规范定义的文档元数据，并将结果存储在与具体系统相关的数据库中。

目前国内外对元数据自动抽取已有不少研究，尤其是国外的一些长期保存项目还开发了相应的元数据自动抽取工具，本指南选取其中几个对其功能、能抽取的元数据信息以及输出格式进行简单介绍，以便在开发国家数字图书馆长期保存系统时参考借鉴。

5.2 DROID （Digital Record Object Identification）

DROID 是 2005 年由英国国家档案馆数字资源长期保存小组开发的，能实现对批量文件格式的自动识别，其目的是满足任何数字知识库准确识别所存储数字对象格式的基本需要。

DRIOD 的主要功能是尽可能准确地识别大量的文件格式和版本信息。当可能实现多个匹配时，例如，一种格式的多个版本包含相同的签名字节序列，所有的匹配和匹配的程度（例如，不定的，确定的）都被列出来。DRIOD 还可以指出文件格式（通过内部签名识别）和文件扩展名之间可能的不匹配。

DRIOD 所能识别的格式范围比其它列出的工具（JHOVE 和新西兰国家图书馆元数据提取工具）都广泛，并能指出持久唯一标识符（PUID），国家档案馆注册中心 PRONOM 能识别的格式都分配了 PUID。然而，它不从文件中提取更多的元数据，也不提取通用元数据（例如，创建日期等）。

保存系统能利用 DRIOD 来提供文件格式识别信息，获取以下 PREMIS 必备语义单元：

- 格式

  格式名称和/或格式注册中心名称和格式注册表

目前，DROID 仅能实现对文件 PUID（PRONOM 唯一标识符）、MIMEType（资源的媒体类型）、Format（格式）、Version（签名版本）、Status（状态说明）、Warning（警告信息）的识别。从识别的结果看，DROID 只能对电子文档的外部特征进行识别，对其内容特征如作者、时间等元数据则无法自动抽取。并且，DROID 不能对所有类型电子文档的外部特征进行识别，如不能识别 RM 格式。DROID 的功能是不断完善和发展的，今后会添加对软件类型、硬件环境、压缩算法和字符编码机制的扩展识别。DROID 的处理结果可选择 XML、CSV 格式存档，还可以进行打印和输出结果预览；中英文文档都能被识别，可满足处理中英文文献的需要。

5.3 NLNZ Metadata Extraction Tool

NLNZ Metadata Extractor（简称 Metadata Extractor）是由 SytecResources 为新西兰国家图书馆开发的，主要用于处理数字化文档和提取元数据信息。它能从多种格式中提取元数据，包括 TIFF、JPEG、GIF、BMP、WAV、MPD、HTML、PDF、MS Word 2、MS Word 6、Word Perfect、MS Excel, MS PowerPoint、MS Works、Open Office。它可以通过一个特定的模块“适配器”提取元数据，并能以“本地适配器”模式或遵守新西兰国家图书馆元数据保存的模式输出 XML。

其可抽取的元数据信息主要为：元数据项（文件名、URL、URI、文件类型、修改时间等）、文件类型元数据（软件相关 ID 信息、开发商、版本、加密算法等相关信息）。

该工具抽取的结果以 XML 形式保存，能直接导入到元数据保存仓储和机构知识库中。该抽取工具从电子文档的头文件中抽取元数据信息，不能对电子文档全文进行抽取。抽取的

字段大都是电子文档的外部特征，如文档名称、类型、修改时间、URL、软件版本等，重要的内容特征如题名、作者、文摘、引文等字段还不能抽取；并且对中文文献不能进行有效的元数据抽取，文件名中中文字符无法显示，只能显示英文和数字部分。

5.4 Matadata Miner Catalogue PRO

Metadata Miner Catalogue PRO（简称 Catalogue）是由 Soft Experience 开发的商业软件，主要可用于抽取题名、作者、主题、关键词等描述性元数据信息。Catalogue 可从 Microsoft Office、Open Office、Star Office、Visiodocument、HTML、PDF、JPEG、TIFF、PSD 文档中实现元数据的自动抽取。

针对不同格式的文档，该工具主要可抽取的元数据信息为：文档类别、题名、作者、主题、关键词、页数、段落数、行数、创作修改时间、PDF 文档的版本、图片编辑状态等

Catalogue 对整个文件夹或多个文档进行识别，自动抽取出元数据信息，并可对自动生成元数据进行修改和补充。在抽取元数据前，用户可自定义需抽取元数据的字段。Catalogue 提供 HTML、CSV、Word、XML 格式的元数据报告，以后还可以生成 Excel 报告。XML 格式的元数据报告可直接用于数据交换和共享，还可以用 XML 专业工具将 XML 输出文档整合到元数据数据库中。

不过该工具对中文文献则不能实现元数据的正常抽取。

5.5 JHOVE（JSTOR/Harvard Object Validation Environment）

JHOVE 是 JSTOR 和哈福大学联合开发的一种用于对象验证的可扩展框架（JSTOR/Harvard Object Validation Environment，JHOVE），它最初设计的目的是识别多种格式，并根据它们声称的格式验证文件。它也能识别格式子类型和版本。在特性描述文件中，JHOVE 也能从多种格式中提取技术元数据，并输出 XML 编码或简单文本。

目前，它的模块能描述 12 种主要的格式类型，包括这些格式的大约 52 个版本或不同的子类型。目前能识别和提取元数据的格式范围包括：TIFF（包括 DNG）、JPEG、JPEG200、GIF、WAV（包括 BWF）、AIFF、HTML、XML、ASCII、UTF-8、PDF（包括 PDF/A）、“字节流”。如果某种格式无法识别，它被归入“字节流”类，并且总合法有效。

它能提取的元数据相当广泛。对于图像和音频，该工具能输出 XML，静态图像转化成 MIX 模式，音频对象和时间码格式转化成音响技术工作者协会（AES）模式。

5.6 建议

从上述介绍的几个工具可以看出，国内外对元数据抽取技术的研究和探索，为进行元数据自动抽取实践提供了一定的技术支持和理论依据。然而目前元数据自动抽取技术也还存在着很大的局限性，还有许多问题需要解决，例如电子文档的格式有所限定，元数据自动抽取

工具对内容元数据的抽取效果欠佳，未实现与数据库系统的有效集成，中文文献元数据自动抽取研究尤其有待提高。国家图书馆在开发长期保存系统的过程中需将现有的元数据自动抽取工具进行汉化，提高其处理中文文献的能力；并在此基础上进行二次开发，设计与具体数据库系统交互的开放接口，实现元数据自动生成并导入相关数据库，才能真正满足国家数字图书馆长期保存的需求。

6 虚拟情景应用实例

保存元数据的应用场景包括：

- 提供长期存取－允许在未来的某一时间点，数字对象能被检索和发现。
- 支持数字对象的存取、传递、显示和执行，并保障传递的内容被读者解读和理解。
- 证明数字对象的真实性，并记录数字对象变化的历史。
- 监测并发现存在风险的数字对象，以采取保护措施。
- 支持数字保存系统的规划和管理，例如，评估在特定的数字对象集上执行特定任务所需的资源（例如，时间，存储容量）。
- 能恢复和重建数字对象，例如，在转换过程发生若干年后才发现存在错误
- 能够向另外的保存系统交接保管权并传递一个、多个或全部的数字对象。

这些假想的场景看似简单，但却足以代表复杂的可能发生的情况。因为无法预见 10 年后会发生什么情况，更别说长期保存了。因此，明智的做法是收集尽可能多的元数据“以防万一”，因为当未来出现问题时，可能无法及时获取元数据。

PREMIS 数据字典对于记录事件有如下阐述：“事件是一个动作，该动作涉及至少一个保存系统已知的对象或代理。”“修改（也就是，创建一个新版本）数字对象的动作文件对于维护履历信息至关重要，是真实性的关键元素。”“即使动作没有做出任何改变，例如对象的有效性和完整性检查，也要重点加以记录，以方便管理。”

这些要求主要涉及上述环境下的事件，即与数字对象“原版”或存档用副本（"master" or archival copy）的保存相关的动作。人们认为保存系统通常除了保存外还有其它目的，而且除了存档“内容”数字对象，支持文件、元数据等的展示文档也以数字对象的形式存放于保存系统中。保存系统记录各种目的的动作和事件。

针对国家图书馆各类型的数字资源以及长期保存各环节的管理需求，本指南分别设计了几个实用情景，以便检测《国家数字图书馆长期保存元数据规范》的科学性、合理性与实用性，同时演示长期保存元数据规范在长期保存系统中可能的记录形式。实例的取值都是示例性的，没有真实意义。

虚拟情景 1：一个数字对象上执行的、不改变这个对象的动作

这个实例适用于如下动作，例如，PREMIS 的事件类型

- 数字签名确认
- 电文摘要计算（校验和计算）
- 固定性检查（校验和检查）
- 确认（格式，例如，一个文件是否符合它的文件名所隐含的格式规范）
- 病毒检查

电文摘要计算和格式确认最好在数字对象摄取时完成，如果可能的话，固定性检查和病毒检查也应于数字对象摄取时完成。在这种情况下，它们可能会，也可能不会被记录为单独的事件，但如果没被记录的话，它们应该在事件详细信息或保存策略和程序中予以指出。如果这些动作没在摄取时完成，它们可能会过一段时间，当能记录为单独的事件再执行。

具体来说，假定要将国家图书馆数字化的地方志 XX 摄入保存系统，以其中一个完整的 PDF 图像文件为对象。保存系统收割前需要对其进行数字签名确认、电文摘要计算、固定性检查、格式确认、病毒检查等，还需要与其对应的 XML 文本文件进行关联。这个过程需要记录的保存元数据具体如下表：

语义单元			受控词表/要遵循的取值规范	语义单元值
对象标识符	对象标识符类型		对象唯一标识符命名域	
	对象标识符值			PDO0000001
对象类型				文件
保存级别	保存级别值		保存系统定的义保存级别值的受控词表	完全保存
	保存级别职责		保存系统定义的保存级别职责的受控词表	要求
	保存级别指定日期		GB/T7408	2010-01-01T01:01:01+8:00
重要属性	重要属性类型			内容
	重要属性值			所有的文字内容和图像
对象特征	组分级别			0
	固定性	电文摘要算法	保存系统定义的电文摘要算法受控词表	MD5
		电文摘要		7c9b35da4f2ebd436f1cf88e5a39b3a257edf4a22be3c955ac49da2e2107b67a1924419563
		电文摘要		保存系统

		创建者			
	大小				2038937
	格式	格式标记	格式名称	保存系统定义的格式名称受控词表	TIFF/NLC/PDF
			格式版本		
	创建程序	创建程序名称			国图将地方志数字化为 TIFF 格式所用的程序
		创建程序版本			国图将地方志数字化为 TIFF 格式所用程序的版本
		创建日期		GB/T7408	2003-01-01T01:01:01+8:00
原始文件名称					北京地方志 XX
存储	内容位置	内容位置类型		保存系统定义的内容位置类型受控词表	URI
		内容位置值			http://xxx.162.59.44/xx/xx
	存储载体			保存系统定义的存储载体受控词表	硬盘
环境	环境性质			保存系统定义的环境性质受控词表	工作环境
	环境目的			保存系统定义的环境目的受控词表	阅读
	环境附注				该环境假定 PDF 本地存储并使用独立的 PDF 阅读器
	软件环境	软件名称			Adobe Acrobat Reader
		软件版本			6.0
		软件类型		保存系统需给出相应的受控词表	显示软件
		软件其他信息			
		软件附件			
	硬件环境	硬件名称			Intel Pentium III 1 GB DRAM
		硬件类型		保存系统需给出相应的受控词表	处理器
		硬件其他信息			最小 32MB
签名信息	签名编码				

	签名人			国家图书馆
	签名方法			DSA-SHA1
	签名值			juS5RhJ884qoFR 8flVXd/rbrSDVGn 40CapgB7qeQiT +rr0NekEQ6BHh UA8dT3+BCTBU QI0dBjlml9lwzEN XvS83zRECjzXb MRTUtVZiPZG2p qKPnL2YU3A964 5UCjTXU+jgFum v7k78hieAGDzNc i+PQ9KRmm//icT 7JaYztgt4=
	签名确认规则			
	签名属性			
	密码信息			
关系信息	关系类型		保存系统定义的关系类型受控词表	源流
	关联对象标识信息	关联对象标识符类型	数字对象唯一标识符命名域	
		关联对象标识符值		PDO0000002
	关联事件标识信息	关联事件标识符类型	事件唯一标识符命名域	
		关联事件标识符值		PE00001
链接事件标识符	链接事件标识符类型		事件唯一标识符命名域	
	链接事件标识符值			PE0002
	链接事件标识符值			PE0003
	链接事件标识符值			PE0004

	链接事件标识符		PE0005
链接权利声明标识符	链接权利声明标识符类型	权利声明唯一标识符命名域	
	链接权利声明标识符值		PRS0001
事件标识符	事件标识符类型	事件唯一标识符命名域	
	事件标识符值		PE0002
	事件标识符值		PE0003
	事件标识符值		PE0004
	事件标识符值		PE0005
事件类型			收割
事件日期		GB/T7408	20101-01T01:01:01+8:00
事件结果信息	事件结果	保存系统定义的事件结果受控词表	成功
链接代理标识符	链接代理标识符类型	代理唯一标识符命名域	
	链接代理标识符值		PA00001
	链接代理功能	保存系统定义的链接代理功能受控词表	执行程序
代理标识符	代理标识符类型	代理唯一标识符命名域	
	代理标识符值		PA00001
代理名称			国家数字图书馆长期保存系统
代理类型		保存系统定义的代理类型受控词表	软件
权利声明	权利声明标识符 / 权利声明标识符类型	权利声明唯一标识符命名域	
	权利声明标识符 / 权利声明标识符值		PRS0001
	权利基本原则	保存系统定义的权利基本原则控词表	版权

	版权信息	版权状态	保存系统定义的版权信息受控词表	公众领域
		版权管辖区域	ISO3166	cn
		版权状态颁布日期	GB/T7408	2003-01-01T01:01:01+8:00

虚拟情景 2：一个数字对象上执行的、把这个对象变成一个新的对象的动作

这个实例适用于如下动作，例如，PREMIS 的事件类型

- 迁移（成为另一种格式）
- 标准化（变成标准或支持的格式）
- 压缩

所有改变数字对象的保存拷贝的事件，都应该被记录。

具体来说，将国家数字图书馆保存系统中的民国期刊 XX 的图像格式由 TIFF 转换为 JPG，这个过程需要记录的保存元数据具体如下表：

语义单元				受控词表/要遵循的取值规范	语义单元值
对象标识符	对象标识符类型			对象唯一标识符命名域	
	对象标识符值				PDO0000005
对象类型					文件
保存级别	保存级别值			保存系统需定义保存级别值的受控词表	完全保存
	保存级别职责			保存系统需定义保存级别职责的受控词表	要求
	保存级别指定日期			GB/T7408	2010-01-01T01:01:01+8:00
对象特征	大小				1024
	格式	格式标记	格式名称	保存系统定义的格式名称受控词表	JPG
			格式版本		2000
	创建	创建程序名称			国图将 TIFF 格式转为 JPG 所用的程序

	程序	创建程序版本		国图将 TIFF 格式转为 JPG 所用程序的版本
		创建日期	GB/T7408	2010-01-01T01:01:01+8:00
原始文件名称				民国期刊 XX
存储	内容位置	内容位置类型	保存系统定义的内容位置类型受控词表	URI
		内容位置值		http://xxx.162.59.44/xx/xx
	存储载体		保存系统定义的存储载体受控词表	硬盘
环境	环境性质		保存系统定义的环境性质受控词表	工作环境
	环境目的		保存系统定义的环境目的受控词表	转换
关系信息	关系类型		保存系统定义的关系类型受控词表	源流
	关联对象标识信息	关联对象标识符类型	数字对象唯一标识符命名域	
		关联对象标识符值		PDO00000004
事件标识符	事件标识符类型		事件唯一标识符命名域	
	事件标识符值			PE00003
事件类型				转换
事件日期			GB/T7408	20101-01T01:01:01+8:00
事件结果信息	事件结果		保存系统定义的事件结果受控词表	成功

虚拟情景 3：从保存系统中删除对象

保存系统对于哪些地方允许删除对象，拥有自己的策略。即使数字对象本身被删除，有关这个对象的一些元数据也应该保留。至少应该保存对象标识符和一些描述性元数据（或一个到描述性元数据的链接，例如通过与当前对象的关系）。

假定国家数字图书馆根据保存策略，在 10 年后要删除保存系统中地方志 XX 之前的老版本，这个过程中应记录的保存元数据具体如下：

语义单元	受控词表/要遵循的	语义单元值

			取值规范	
对象标识符	对象标识符类型		对象唯一标识符命名域	
	对象标识符值			PDO0000007
对象类型				文件
保存级别	保存级别值		保存系统定的义保存级别值的受控词表	完全保存
	保存级别职责		保存系统定义的保存级别职责的受控词表	要求
	保存级别指定日期		GB/T7408	2010-01-01T01:01:01+8:00
重要属性	重要属性类型			内容
	重要属性值			所有的文字内容和图像
重要属性	重要属性类型			页数
	重要属性值			300
对象特征	组分级别			0
	固定性	电文摘要算法	保存系统定义的电文摘要算法受控词表	MD5
		电文摘要		7c9b35da4f2ebd436f1cf88e5a39b3a257edf4a22be3c955ac49da2e2107b67a1924419563
		电文摘要创建者		保存系统
	大小			2038937
	格式	格式标记 格式名称	保存系统定义的格式名称受控词表	JPG
		格式标记 格式版本		2003
		格式注册中心 格式注册中心名称		国家数字图书馆

	创建程序	创建程序名称		国图将地方志数字化为 JPG 格式所用的程序
		创建程序版本		国图将地方志数字化为 JPG 格式所用程序的版本
		创建日期	GB/T7408	2003-01-01T01:01:01+8:00
原始文件名称				北京地方志 XX
存储	内容位置	内容位置类型	保存系统定义的内容位置类型受控词表	URI
		内容位置值		http://xxx.162.59.44/xx/xx
	存储载体		保存系统定义的存储载体受控词表	硬盘
关系信息	关系类型		保存系统定义的关系类型受控词表	源流
	关联对象标识信息	关联对象标识符类型	数字对象唯一标识符命名域	
		关联对象标识符值		PDO0000009
	关联事件标识信息	关联事件标识符类型	事件唯一标识符命名域	
		关联事件标识符值		PE00006
事件标识符	事件标识符类型		事件唯一标识符命名域	
	事件标识符值			PE0002
事件类型				删除
事件日期			GB/T7408	2011-01-01T01:01:01+8:00
事件结果信息	事件结果		保存系统定义的事件结果受控词表	成功
链接代理标识符	链接代理标识符类型		代理唯一标识符命名域	
	链接代理标识符值			PA00001
	链接代理功能		保存系统定义的链接代理功能受控词表	执行程序

代理标识符	代理标识符类型	代理唯一标识符命名域	
	代理标识符值		PA00001
代理名称			国家数字图书馆长期保存系统
代理类型		保存系统定义的代理类型受控词表	软件

参考文献

- [1] Data Dictionary section from PREMIS Data Dictionary for Preservation Metadata version 2.1.[2011-09-01].http://www.loc.gov/standards/premis/v2/premis-dd-2-0.pdf
- [2] Implementing Preservation Repositories For Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community——A Report by the PREMIS Working Group September 2004. [2010-3-30]. http://www.oclc.org/research/activities/past/orprojects/pmwg/surveyreport.pdf
- [3] Implementing Preservation Metadata in Digital Library Applications: Using PREMIS with METS. Rebecca Guenther (LC), Tom Habing (UIUC), Nancy Hoebelheinrich (Stanford), Ardys Kozbial (UCSD), Rob Wolfe (MIT) DLF Spring Forum, April 2008. [2009-11-10]. http://www.diglib.org/forums/spring2008/2008springprogram.htm
- [4] Using Metadata Standards in Digital Libraries: Implementing METS, MODS, PREMIS, and MIX. Rebecca Guenther (LC), Morgan Cundiff (LC), Nate Trail (LC), Brian Tingle (CDL), Sarah Shreeves (UIUC), Tom Habing (UIUC), Tod Olson (U Chicago) ALA Annual 2007. [2010-03-30]. http://www.loc.gov/mods/presentations/litaprogram-an2007.html
- [5] Implementing the PREMIS data dictionary: a survey of approaches. [2010-05-10].http://www.loc.gov/standards/premis/implementation-report-woodyard.pdf/
- [6] Australian Partnership for Sustainable Repositories PREMIS Requirement Statement Project Report.[2010-4-12].http://www.apsr.edu.au/publications/presta.pdf
- [7] Understanding PREMIS. [2010-10-29]. http://www.loc.gov/standards/premis/understanding-premis.pdf