网络出版时间：2016-11-04 15:06:08
网络出版地址：http://www.cnki.net/kcms/detail/31.1108.G2.20161104.1506.022.html

以事件为核心的长期保存数字起源管理框架研究

吴振新[1]，李文燕[1,2]（1 中国科学院文献情报中心 2 中国科学院大学）

摘要 首先分析了长期保存领域相关标准 OAIS，PREMIS 和 TRAC 对起源信息的解释和要求，然后对比分析了已有的长期保存系统对起源的应用情况。在此基础上，提出了以事件为核心记录起源的应用框架，并对关键问题加以分析，包括起源的内容，组织方案，存储封装和技术实施方案，为长期保存系统组织管理起源信息提供参考。

关键词 起源 事件 长期保存 管理框架

The Study of Event-Centric Provenance Management Framework within Long Term Preservation

Wu Zhenxin[1] Li Wenyan[1,2]（1 National Science Library，Chinese Academy of Sciences 2 University of Chinese Academy of Sciences）

Abstract Firstly, it analysis the explanations of provenance of the relevant standards ，namely OAIS, PREMIS and TRAC.Secondly, it makes a comparative study of the application in the existing long-term preservation systems. At last it puts forward an event-centric provenance management framework in data preservation. And further more,it summarizes key elements in practice including provenance content, organization scheme, storage and package method , and technical implementation scheme，with aim to provide a reference for data preservation system to manage the provenance.

Key words Provenance, Event, Data preservation for a long time, Management framework

起源，即 Provenance，代表了数字对象的产生及发展历史。通过记录起源信息，人们可以了解数字对象所发生的变化，以及变化的地点、原因、时间和责任人等 7W[1]信息（What、Where、Who、When、Which、Why、How）。起源功能对于解决重要数据问题至关重要：数据的可信性，结果可靠性，数据修改或分析过程的透明性以及对数据引用的支持。

长期保存系统的首要目标是保证起源信息的真实性，可理解性和可访问性，如果数据失去了真实性，可理解和可访问也就无从谈起。数字对象的真实性在保存的过程不是不变的，需要确保数字对象正如其声明的那样。但是数字资源在保存过程因为格式转换、媒体迁移、规范化等原因会产生多种变化。起源信息可以提供真实性判断的证据，显示数字发生了什么变化，对数字对象产生了什么影响，以此证明数字对象是否真实。同时起源信息还可以提供数字版权的证据，验证科学实验数据等。

起源信息的追踪和记录是数字文献资源长期保存机构数据基础设施须提供的一项服务，目的是确保数据真实性、透明性、可追溯性以及可复用性。已有长期保存项目中大多从一开始就注重起源的记录，但是采用的技术、方法存在许多的不同。针对以上问题，本文研究了长期保存相关标准对起源信息的要求和应用实践，全面分析了起源在长期保存系统中的应用，在此基础上，分析并总结并提出了以事件为核心记录起源的应用框架，为长期保存系统组织管理起源信息提供参考。

1 相关标准对起源的要求和描述

数字起源是一种重要的元数据信息，长期保存的相关标准和规范都对其作出了定义和解释，尤其是 OAIS，PREMIS 和 TRAC，分别从定义，描述方法和必要性对其加以描述。

1.1 OAIS 的要求与描述

OAIS 是长期保存的基础框架，它为长期保存提供了标准和规范化的流程和信息对象模型。OAIS 对起源的定义，内容和作用等做了简要陈述。在 OAIS 中，起源被定义为内容信息的历史，其展示了内容信息产生的由来，从产生以后发生的变化，和自创建以后的保管人的变动[2]。例如，数字图书馆集合起源包括却不限于：

——非原始数字内容：包括数字过程和主要版本链接。

——数字出版物：原始版本链接，保存过程的元数据，更早版本的链接，改变历史和信息对象描述。

不同类型的数据的起源数据记录的内容有所不同，如对于空间科学数据，起源需要记录仪器信息，主要研究者，软件接口规范信息等；对于软件包，需要记录修改历史，注册信息和版权等内容。

起源是 PDI 一部分。OAIS 中信息对象包括两大部分，如图 1 所示，信息包由内容信息和保存描述信息（Preservation Description Information，PDI）组成，其中 PDI 主要负责解释内容信息。起源是 PDI 的重要组成，它记录了数字对象版本变化，格式转换，人为失误等内容，为可信赖认证，信息审计，权限判断，版本变迁等提供重要依据。

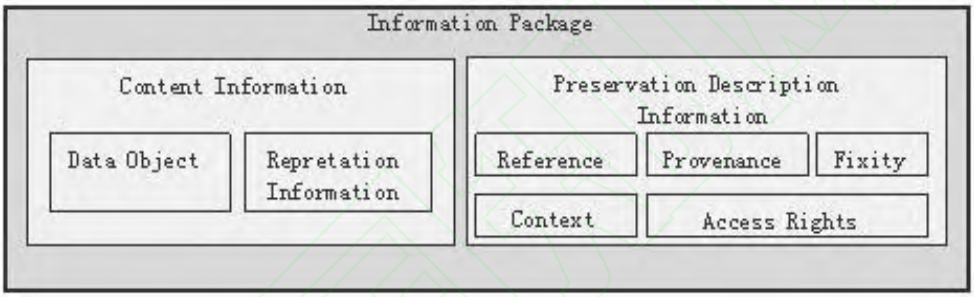

图 1.OAIS 中信息对象概念模型

起源信息记录和维护不仅限于长期保存过程中，它的产生贯穿了整个数字对象生命周期，但在保存过程中，摄入前的起源信息则应由内容生产者提供，保存过程中则应由保存系统负责创建和维护从摄入开始的起源信息。

1.2 PREMIS 要求与描述

PREMIS 是支持数字保存处理过程的信息模型框架，包括 PREMIS 数据字典[3]和 PREMIS 框架[4]。PREMIS 对起源做了更加详细的陈述，并提供了可描述起源的保存元数据。

PREMIS 数据字典定义，起源主要描述了代理对数字对象保管和管理责任，发生在数字对象生命周期内的关键事件，以及其他与数字对象的创建，管理和保存有关的信息。记录起源是保证对象可信赖的重要手段，可以从技术层面为真实性管理提供了支持。

PPREMIS 框架对起源作了更丰富和深刻的阐释。一方面，起源信息主要解决了内容数据对象的时间方面的问题，包括从对象在存档系统中被创建一直到其当前状态。另一方面，起源信息是基于事件的元数据，对象的演化过程可以归结为被重要事件驱动并得以体现的过程，例如对象的创建，被摄入到保存系统，所有权的转移，格式迁移等，无一不伴随着相关事件的发生。如图 2 所示，从完整生命周期来看，起源信息包含来源，摄入前，摄入，存档，权限管理 5 个方面的事件。

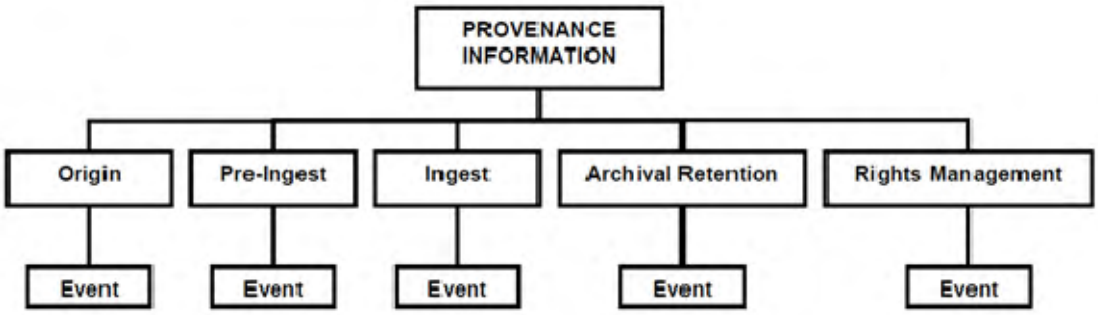

图 2.PREMIS 框架定义起源事件模型

记录特定事件的细节，以及它们对内容数据对象的影响，是起源信息的一个关键功能，而事件是 PREMIS 数据字典 5 大基本实体之一，通过事件语义单元可以描述起源信息。然而事件只是起源的核心部分，从 OAIS 和 PREMIS 给的定义来看，还应该包括操作过程中涉及的对象和责任者。因而，可以从 PREMIS 中抽取出如表 1 所示相关语义单元来描述起源。

表 1 描述起源的 PREMIS 语义单元

语义单元	描述元素
Event	2.1 eventIdentifier 2.1.1 eventIdentifierType 2.1.2 eventIdentifierValue 2.2 eventType 2.3 eventDateTime 2.4 eventDetail 2.5 eventOutcomeInformation 2.5.1 eventOutcome 2.5.2 eventOutcomeDetail 2.5.2.1 eventOutcomeDetailNote 2.5.2.2 eventOutcomeDetailExtension 2.6 linkingAgentIdentifier 2.6.1 linkingAgentIdentifierType 2.6.2 linkingAgentIdentifierValue 2.6.3 linkingAgentRole 2.7 linkingObjectIdentifier 2.7.1 linkingObjectIdentifierType 2.7.2 linkingObjectIdentifierValue 2.7.3 linkingObjectRole
	1.10relationship 1.10.3 relatedObjectIdentification 1.10.3.1 relatedObjectIdentifierType 1.10.3.2

Object	relatedObjectIdentifierValue 1.10.3.3 relatedObjectSequence 1.10.4 relatedEventIdentification 1.10.4.1 relatedEventIdentifierType 1.10.4.2 relatedEventIdentifierValue 1.10.4.3 relatedEventSequence 1.11 linkingEventIdentifier 1.11.1 linkingEventIdentifierType 1.11.2 linkingEventIdentifierValue
Agent	3.1 agentIdentifier 3.1.1 agentIdentifierType 3.1.2 agentIdentifierValue 3.2 agentName 3.3 agentType

eventType 是受控词，每个仓储须定义自己的 eventType 值，这也是记录起源的一个难点和重点。PREMIS 提供了一个事件类型清单供参考：creation、deacession、decompressionn、decryption、deletion、digital signature validation、dissemination、fixity check、ingestion、message digest calculation、migration、normalization、 replication、validation 和 virus check。

1.3 可信赖仓储认证标准的要求与描述

可信赖仓储认证标准（TRAC）即 ISO1636 为数字仓储库的真实性的审核和认证提供基础框架，是数字仓储库真实性审查的标准规范。

TRAC 中，对起源的描述主要出现第 4 章和第 5 章，解释了起源发挥的作用，必要性和维护要求等。

起源提供复制和移动数据的过程信息，且必须被不断维护和升级，可以帮助确定责任人，数字对象副本的数量和位置。当 SIP 与 AIP 出现不一致时，仓储须根据书面规程进行处理，并且需要指明不一致的原因，起源可以发挥重要的作用。PDI 通过提供起源信息以及与其他信息之间的关联，确保内容信息能够被理解，这也是理解内容信息的关键元素。

根据协议，在数据对象处理过程中，除非协议另有说明，仓储可通过文档格式来判断保存对象的相关属性。在这种情况下，仓储需要对格式相同的保存对象的起源信息进行统一描述。为了使仓储须拥有一套能够支持长期保存的 AIP 定义，必须能够识别和解析 AIP 中的必要组件。

因此，保存仓储需要有文档清晰地展示诸如表征信息和起源信息之类的 AIP 组件，使之能够被管理和及时更新。而且起源信息还要和 AIP 的关键信息如内容信息，表征信息和其他 PDI 关联起来，并对它们之间的关联有一个统一的定义。为了有效识别和解析起源，仓储还需要拥有一套机制来正确验证所有内容生产方的身份信息，支持长期保存的 AIP 定义，并根据实际情况随着时间扩展起源。

1.4 小结

通过对以上三个标准的分析，我们初步了解了起源信息的内涵，构成和作用，为我们在实践中管理起源提供了指导，说明和规范。OAIS 对起源信息的概念，以及与内容信息和其他 PDI 信息的关系作出解释，是长期保存起源管理实践首先要参考的标准。PREMIS 则针对起源的内涵作了更深入的说明，指出起源展现数字对象变化的能力是事件驱动的，因此可以通过记录事件来组织起源信息，同时提供了表 1 所示的以事件为主的语义单元组织起源信息，具有重要实践意义。而根据 TARC，我们知道起源是构建可信赖仓储的重要证据，不仅需要记录起源信本身，而且还要维护起源信息和内容信息以及其他 PDI 信息之间的联系，同时因为起源是随时间而不断增长的，所以需对其不断的更新和维护，这样才能满足可信赖仓储认证的整体要求。

2 起源在长期保存中的应用现状

在第一版 OAIS 中，已规定起源是 PDI 的一部分，但并未指出具体使用何种元素和方法记录起源信息，在长期保存中对起源的研究相对较广，许多项目都对其展开了探索，深度和方式各有不同。

2.1 DAITSS

DAITSS[5] 是由佛罗里达图书馆自动化中心为佛罗里达数字保存系统(FloridaDigitalArchive, FDA)开发的一个数字保存仓储系统。它利用 METS 格式，把起源记录在管理元数据 amdSec 的 digiprovMD 元素里。对于起源，DAITSS 主要以事件来记录，并把事件分为两个级别，包级和文件级。包级事件包括 submit，ingest，disseminate，refresh 和 withdraw，文件级事件包括 virus check，describe，xml resolution，normalize 和 migrate。一个 AIP 的 METS 文件封装了三个级别的管理元数据，分别包含不同层次的起源信息。第一级起源记录了协议信息。第二级起源记录了 PREMIS 提交，摄入，分发，更新和撤销事件。第三级起源记录了 DAITSS 服务为每个文件执行的 PREMIS 事件，如文件转换，病毒检查。

2.2 CASPAR

CASPAR[6]使用 IBM 开发的 PDS[7]来管理起源数据，主要用于权限管理，知识库更新跟踪，和真实性管理。在 PDS 中，起源被当作独立的信息对象处理，拥有自己的表征信息。一条起源记录就是一个起源事件，事件可分为内部和外部两种。起源数据的概念结构如下，其中 PDS internal 表明该起源事件发生 PDS 内部/外部的标志，Content 记录了起源事件的详细信息。一个起源事件可能指向单个 AIP（如创建），或者一组 AIP（如包含某种数据的所有 AIP 被转换成某种新的格式），或者整个系统（如所有存档的所有者发生改变）。

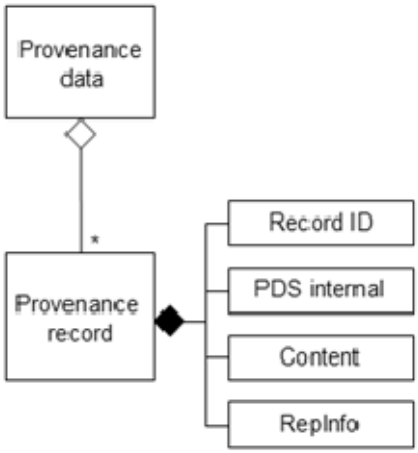

图 3.CASPAR 起源信息概念结构

2.3 APARSEN

APARSEN[8]把起源作为真实性管理的主要证据来来收集，并为此做了一份详细的指导，以及实证来证明其可行性。数字资源的生命周期被分为摄入前和保存期两个阶段，并通过事件记录起源。由于长期保存系统的复杂性，APARSEN 把核心事件整理分类，而不一一罗列。每个阶段包括如下事件类型：

1. 摄入前——捕获，整合，聚合，删除，迁移，转换，提交。
2. 保存期——捕获，保存—摄入，保存—聚合，保存—抽取，保存—迁移，保存——删除，保存—转移。

事件用以下元素加以描述：事件的描述、代理、输入、输出以及可信性证据记录（AER）。AER 指明了应控制和收集那些和真实性和起源相关的信息。

APARSEN 使用 CRM_{dig} 模型描述起源，并对 CRM_{dig} 和 OPM 做出了映射，增加其交互性。CRM_{dig} 是一个以事件为核心来描述起源信息的本体模型，重点突出了对物理对象的数字起源的描述，以及对参与数字化过程中的设备信息的陈述。

2.4 SCAPE

SCAPE[9]的把起源应用到了应用到了 SCAPE 保存规划，数据出版平台和知识库模块。捕获起源方面，起源组件利用 Taverna 插件 Workbench 2.4 输出 Taverna 工作流的起源。Taverna 拥有自己的起源本体—— tavernaprov，该本体扩展了 PROV-O 和 wfprov，目的是描述 Taverna 的特定行为，不包括一般的模型，如错误文档和迭代。在 SCAPE 信息包中，使用了 METS 文件的 digiprovMD 元素封装起源[10]，起源记录由 PREMIS 事件和代理组成，利用 premis:object，premis:event 和 premis:agent 相关语义单元描述起源。

2.5 其他相关项目

长期保存的其他许多项目和系统也都涉及起源的收集和管理。IRODS[11]设计了分布式的起源信息系统，提供多结点的起源记录(“P-Services”)和起源查询(Q-Services)服务.记录的起源事件表示规则的执行，不仅包括内容数据和文件的变化历史，而且记录用户对文件访问，处理数据的规则版本变化和IRODS的系统信息等。PrestoPrime[12]通过事件和责任人如生产者来记录起源，事件被划分为存缴前事件和存缴事件，后者包括新版本产生和有效性

检查，参考PREMIS字典，使用DNX和OPM两个模型来记录起源。Data Conservancy[13]把起源划分为起源服务和世系服务两部分，前者记录了发生在系统内的事件，后者记录了数据对象之间的关系，并通过HTTP调用Lineage API和Event API两个web接口来调用上述服务。

2.6 项目小结

起源的管理关键在于设计起源组织模型和记录流程。在对起源的组织方面，虽然不同保存系统使用了不同组织模型如 OPM，PREMIS 和 CRM_{dig} 或者自定义，但是却不约而同的以事件为核心来记录起源，这一点和 PREMIS 是一致的。除事件之外，还记录了生产者和对象关系这些重要信息。在对起源进行管理时，把起源信息作为元数据和内容信息一起保存，并按照一定的封装格式如 METS 进行组织。在起源组织管理方面，和技术元数据以及描述元数据不同，起源信息在需要不断的更新，一条起源信息一般需要经过捕获，组织，封装存储这样的过程，并最终提供访问查询，或者被保存系统的其他模块调用。

3 基于 OAIS 框架的起源信息管理框架

基于以上分析，本文总结并设计了以事件为核心的起源信息管理框架。对 PREMIS 提供了 15 个事件类型，加以扩展。综合参考 PREMIS 事件清单、TRAC 和 OAIS 要求以及相关保存系统实践，列出典型的保存事件。如图 3 所示，起源事件贯穿了从摄入开始至对象消亡整个长期保存过程，每个过程也所包含不同类型的事件，其中主要的事件发生在摄入和存储过程中。

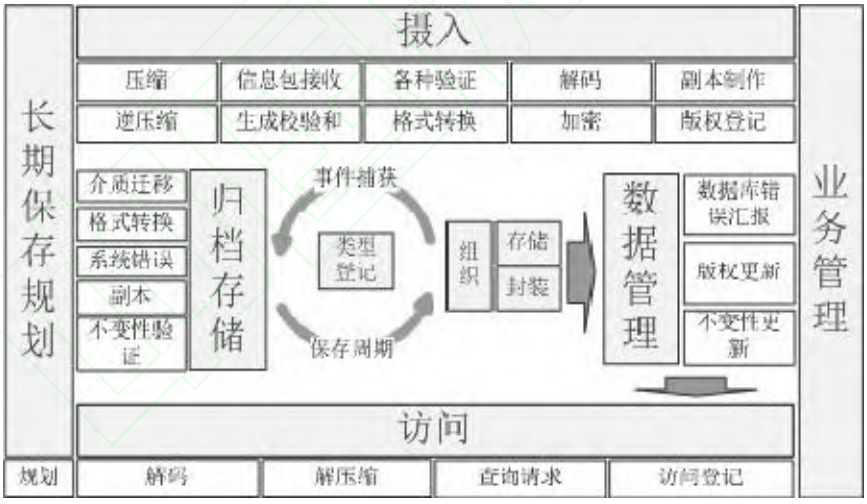

图 4.长期保存起源管理框架

以事件为核心的起源的管理共涉及了四个基本流程模块，即捕获，组织，封装存储和访问，其中捕获，组织以及存储封装是重点。捕获指的是根据事先设定好的事件类型，监控保存系统的相关内容，一旦事件发生，将其记录下来，并通知组织模块。组织模块则按照设计好的的起源模型和元素，以规范的方式把起源记录下来产生相关的文件。存储封装则负责则起源信息存储到相关的信息包文件或者数据库中，并对起源文件按照一定的格式封装保持其与数字对象之间的关联。访问主要提供用户查看和下载起源信息以及应用嵌入到长期保存信息其他功能中，如真实性管理。

4 数字起源管理框架的关键问题分析

以事件为核心的起源信息管理框架解决了起源信息组织和应用的一般管理流程，接下来本文就框架涉中及到的关键问题，基于整个保存周期对起源的内容、组织方案、存储封装和技术方案进行全面的探索。

4.1 起源信息的内容

起源信息指的是数字对象及变化的信息，在长期保存系统中，以事件为核心的起源主要记录了以下相关内容：

（1）发生数字对象上的事件

通过事件的连接，数字变化得以呈现，事件的细节描述是起源信息的重点内容，除了包含事件标识符，细节描述，时间，责任人和涉及对象基本信息，还应该包含事件类型，处理设备，处理结果，发生原因等内容。

（2）关联的数字对象

应该完整保存事件发生过程涉及的对象的关联，否则起源变得毫无意义。有时，一条起源同时关联两个甚至更多对象，如某事件产生了一个新的对象，则该起源信息同时术语这两个对象。

（3）代理内容

狭义的代理主要指的操作人，长期保存系统有责任记录相关责任人在其中发挥的作用，这些操作是否合法。但是更广义的代理在此指的不仅是人，它包括组织，个人或者软件产品等。

（3）对象之间的关系

对象在提交到保存系统后，常常有多个副本或者不同格式的版本。在向用户发送 DIP 时，需要对其加以说明，原始版本是否发生变化，如果变化，提供现存版本，并加以说明。

4.2 起源信息的组织方案

在《起源信息模型及标准 PROV 的研究分析》[14]中作者对起源信息的模型做了详细的论述，在此，本文只简要分析几种重要的起源模型。起源的组织研究主要分为两个方面，即通用描述模型和元数据或本体词汇。

通用描述模型，典型的有 W3C PROV，OPM 和 Provenir 等。虽然模型之间对起源的元素定义差异较大，但在对起源的基本元素却基本是一致的，即可识别的对象、处理对象的过程和涉及到的责任方。以 W3C PROV[15]为例，起源信息被定义为一条记录，它描述了在生产、影响、传递对象过程中所涉及的人、机构、实体和活动。PROV-DM 的核心结构定义为 Entity，Activity 和 Agent 及其之间的关系，如图 5 所示。映射到长期保存中，Entity 对应保存对象，Agent 责任人和处理设备，Activity 对应起源事件。

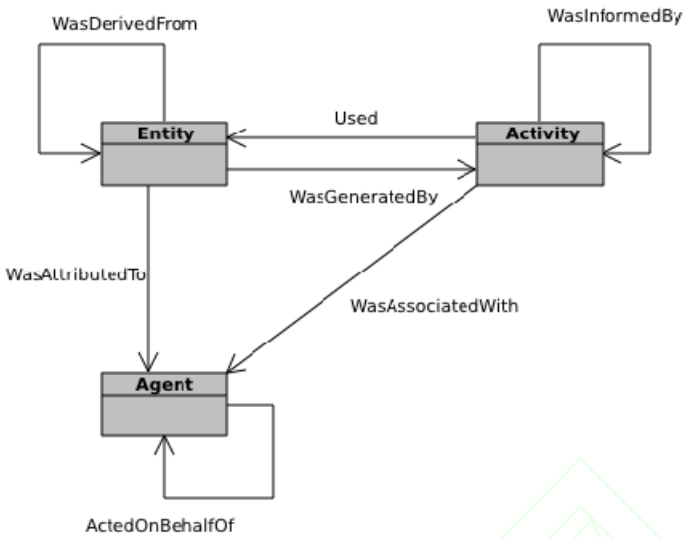

图 5. PROV 核心结构

元数据方案、本体词汇有 DC 元数据、VoID 词汇和 Provence Vocabulary 等，有些虽然不是专门为表达起源而设计的，但是也在一定程度上展示了起源的内容。

起源的组织模型是起源管理的核心，长期保存应用中，起源的组织方式比较多样。较早的做法，如 TNA[16]（英国国家归档中心）自定义元数据，把保管历史，保管人，保存历史事件等字段映射为 Provenance，起源信息分散至多个不同的表结构中，不利于整合。后来较为通用的做法是直接使用 PREMIS 事件实体作为起源信息，例如 DAITSS 和大英图书馆的数字图书馆项目。同期，CASPAR 项目则使用 CRM_{dig} 本来组织起源信息。为了增强起源的交互性和表述能力，如 PrstoPrime 开始使用通用模型 OPM 和 PREMIS 两种方式采集不同的起源信息。近期，SCAPE 则依赖 taverna 的起源插件[17]，使用扩展的 W3C 标准 PROV 模型记录。由此看出长期保存社区对起源信息日益的重视，在组织起源方面，语义和结构越来越丰富,通用交互性和通用性越来越强。

4.3 起源信息的存储与封装

起源的存储方式可以分为两种，一种把起源和其他元数据保存在一起，例如，在 METS 文件中起源信息和版权信息，通过引用外部起源文件或者直接记录起源元数据的方式被保存在管理元数据区域，这种混合存储的方法，优点是易于维护起源的完整性，缺点是难于发布和搜索起源。

另一种，把起源信息单独存储到一个起源仓储中。DataONE[18]把工作流起源存储在 Mysql 和图数据库构建的起源仓储中。这种方式，便于起源的快速查询和可视化呈现，缺点是维护困难，需要考虑是起源一致性，当数据被修改时起源版本是否需要变更等问题。这两种方式常常综合使用，把经常访问的起源信息和其他元数据一起存储，需要时抽取出来，经常使用的起源则存储到单独的文件或数据库系统中，用以提供快速访问。

相应的起源的存储方式也有文件，如如 XML，RDF 和 OWL 或数据两种。在以文件方式存储时常常涉及起源信息的封装，常使用的封装方式有 METS 和 XFDU。

4.3.1 DAITSS 与 METS

METS[19]由美国数字图书馆联盟 DLF 开发,用来将一个数字对象及其相关的描述性元数据、管理性元数据和结构性元数据进行封装和编码的标准规范，应用较广泛，DAITSS 和 UK 期刊保存等项目使用都是这种方法。

METS 文档主要包括 7 个部分，在 DAITSS 保存系统中，起源信息被封装在管理元数据管理元数据 amdSec 部分的 digiprovMD 元素里。DAITSS 项目 SIP，AIP 和 DIP 分别保存不同类型的起源信息。

SIP 中，METS 文件包含有效的 DAITSS 账号代码和有效的 DAITSS 项目代码，并在 <amdSec> 中把两者关联起来，编码实例如下。

```
<METS:amdSec>
    <METS:digiprovMD ID=”[unique id]”>
        <METS:mdWrap MDTYPE=”OTHER” OTHERMDTYPE=”DAITSS”>
            <METS:xmlData>
                <daitss:daitss>
                    <daitss:AGREEMENT_INFO
                    ACCOUNT=”[account code]” PROJECT=”[project code]”>
                </daitss:daitss>
          <METS:xmlData>
              </METS:mdWrap>
       </METS:digiprovMD>
</METS:amdSec>
```

AIP中，一个AIP的METS文件封装了三个不同层次的起源信息：

（1）第一级起源记录协议信息（账户或项目）。

（2）第二级起源记录 DAITSS 信息包级别的事件，包括：PREMIS 提交事件，PREMIS 摄入事件，PREMIS 分发事件，PREMIS 更新事件和 PREMIS 撤销事件

（3）第三级起源记录 PREMIS 事件和代理，事件是指针对文件的 DAITSS 系统服务：

- SIP 的描述文件或其他 XML 文件的 DAITSS 解析服务事件
- DAITSS 转换事件（规范化或迁移）
- 每个提交文件的病毒检查事件

DIP 中， METS 文件记录了文件不同版本之间的联系（即起源），标识出原始提交版本。如果发生变化，则提供最新，最好的版本。一个包含起源的 METS 文件如下

```
<amdSec> <!-- Package-level metadata -->
<techMD ID=”tech-1” ADMID=”dmd-1 digiprov-1 digiprov-2 digiprov-3 digiprov-4”>
<!--Intellectual Entity-->
<techMD ID=”tech-2”> <! -- PREMIS current representation -->
<techMD ID=”tech-3”> <! -- PREMIS normalized representation-->
<techMD ID=”tech-4”> <! -- PREMIS original representation -->
<!--DIGIPROV (events) -->
<digprovMD ID=”digiprov-1”> <!--PREMIS package submit event-->
<digprovMD ID=”digiprov-2”> <!--PREMIS package ingest event-->
<!--DIGIPROV (agents) -->
<digiprovMD ID=”digiprov-3”> <!--DAITSS account agent -->
<digiprovMD ID=”digiprov-4”> <!--DAITSS software agent -->
```

</amdSec>

4.3.2 CASPAR 与 XFDU

XFDU [20]是由美国空间数字系统咨询委员会(CCSDS) 开发的推荐标准，用于打包封装数据和元数据并形成单个包。XFDU 是一个物理容器，由一个 Manifest 文档以及该文档调用文件组成，CASPAR 项目使用这种格式封装起源[21]。

manifest 是一个 XML 文档，包含 packageHeader，dataObjectSection，metadataSection，informationPackageMap 和 behaviorSectioon 五部分。CASPAR 通过两种方式把起源封装在 metadataSection 位置，第一种直接写入，第二种是通过 URL 链接，指向外部的 PDI 文件中的起源信息。在 informationPackageMap 中通过 XML 标识符把内容对象和起源关联起来。

XFDU 严格遵守 OAIS 对信息对象的定义，起源在 xml Schema 记录为 PROVENANCE，类属于 PDI，示例如下：

```
<informationPackageMap ID="informationPackageMap">
<xfdu:contentUnit ID="cu5" order = "1.3" pdiID = "provenance" textInfo = "content unit for hdfFile2" dmdID = "ECSDMD">
 <dataObjectPointer dataObjectID = "hdfFile2"/>
 </xfdu:contentUnit>
metadataSection
</informationPackageMap>

<metadataObject ID = "provenance" classification = "PROVENANCE" category = "PDI">
 <metadataReference vocabularyName = "OTHER" mimeType = "text/xml" textInfo = "processing history XML file"
 locatorType= "URL" href = "file:packagesamples/scenario1/pdi.xml" />
 </metadataObject>
```

4.4 技术实施方案

起源的技术实施方案可以分为以下 4 种类型：

（1）直接开发相关函数模块，接口，或组件。如 CASPAR 项目的 PDS，PDS 主要负责 AIP 存储，并提供感应和记录所有发生 PDS 内的所有事件的功能，同时提供方法帮助用户记录 PDS 无法感应的外部事件。

（2）使用元数据工具抽取起源信息。此方法常常能够把起源抽取为 PREMIS 事件，通过抽取保存系统中对象经历的部分事件，如 JHOVE，DROID 等，目前较为常见。虽然，也有专门针对起源的元数据抽取工具，例如由 ExLibris[22]针对起源开发的元数据工具，遗憾的是此软件并不开源。

（3）使用工作流引擎自带的起源信息记录插件。SCAPE 使用了 Taverna 工作流软件，执行转换，迁移，副本制作等任务，抽取出工作流中的起源信息。

（4）嵌入已有的起源记录引擎。此方法可利用已有的接口，在长期保存系统中集成起源管理的功能。通用的起源引擎有 PASOA 和 KARMA，它们可以和相应的工作流对接，通过基于 web 服务的接口提供起源的记录，存储和查询，目前 IRODS 已实现和 PASOA 的整合。

每种方案适用不同的情况，方案（1）定制开发，比较灵活，可以根据实际需求设计起源组织模型，并编码开发，但相对工作量较大；方案（2）工作量较小，便于集成，但缺点

的是目前已有的元数据抽取工具只能记录部分事件，无法满足长期保存系统的记录完整起源的需求；方案（3）依赖于特定的工作流引擎，系统需要事先集成了特定的工作流引擎，才能使用其插件抽取出起源信息，但相对工作量较小，并支持通用的起源模型如 W3C PROV；方案（4）相对技术依赖性较弱，只要满足引擎的环境需求，即可集成，但需要根据实际要求对起源引擎的起源组织模型进行相关的修改。

5 结语

以上对长期保存领域的数字对象的起源做了比较全面而深入的分析，但实践过程中的复杂环境和多样化的应用需求，还有一些问题需要深入考虑。

数字起源包括的范围十分广泛，并且与 PDI 的其他部分，如情境信息（context）有所交叉，所以在在设计长期保存系统中起源的记录时，应该明确的范围界定，以免发生混淆。

相对于起源的组织、捕获和存储来说，起源信息可视化方法和技术相对比较缺乏，有些系统至今只提供 xml 等方式文件来访问起源，这种不友好的呈现方式既不利于其阅读，也不利于展现起源应有的价值，所以如何有效的多角度、生动化、清晰化的呈现起源信息，是值得长期保存系统关注的事情。

同时，和内容数据一样，起源数据应该被妥善存储和保护。保存机构应该采取措施保护起源的安全性和真实性，在突发事件如数据变换、存档和转换过程使用防篡改技术（如数字签名）以保护起源信息链的完整性、可靠性和有效性。

总的来说，起源对于维护数字对象真实性、版权归属、访问权限、知识库变迁等具有重要作用，它既是 OAIS 信息模型的一部分，又是长期保存系统实践中非常重要的支撑内容，所以长期保存系统应该充分的结合 OAIS、PREMIS 和 TRAC 等标准，根据自身的实际情况，制定出一套完善的起源应用管理方案。

参考文献

[1]Ram S, Liu J. A New Perspective on Semantics of Data Provenance[C]. SWPM, 2009.

[2]CCSDS 650.0-M-2, Reference Model For An Open Archival Information System (OAIS)[S].

[3]PREMIS Editorial Committee. PREMIS data dictionary for preservation metadata, version 2.0[J]. Retrieved November, 2008, 15: 2009.

[4]Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects, a Report[M]. OCLC/RLG Working Group on Preservation Metadata, 2002.

[5] The Florida Center for Library Automation. DAITSS website[EB/OL]. [2015-03-15].http://daitss.fcla.edu/.

[6]Factor M, Henis E, Naor D, et al. Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage[C]//Workshop on the Theory and Practice of Provenance. 2009.

[7]IBM. Preservation DataStore Interface[EB/OL].

[2015-03-15].http://www.casparpreserves.eu/Members/cclrc/Deliverables/updated-preservation-datastores-interface/at_download/file.pdf.

[8] Salza S, Guercio M, Grossi M, et al. D24. 1 Report on authenticity and plan for interoperable authenticity evaluation system[R]. Tech. rep, 2012.

[9]SCAPE website[EB/OL].[2015-03-15].http://www.scape-project.eu/.

[10] Withers D, Paton N. Design of provenance [EB/OL].

[2015-03-15].http://www.scape-project.eu/deliverable/d7-1-design-of-provenance-component.

[11] Weise A, Hasan A, Hedges M, et al. Managing provenance in iRODS[M]//Computational Science–ICCS 2009. Springer Berlin Heidelberg, 2009: 667-676.

[12] Kashi N, Sherwinter N. AV Data Model: Final Specification [EB/OL].
[2015-03-15].https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.3_AV_Data_Model_R0_v1.00.pdf.
[13] Mayernik M S,DiLauro T, Duerr R, et al. Data Conservancy Provenance, Context, and Lineage Services: Key Components for Data Preservation and Curation[J]. Data Science Journal, 2013, 12(0): 158-171.
[14]李文燕,吴振新. 修改稿-起源信息模型及标准 PROV 的研究分析[J].情报理论与实践,2015(04).
[15]W3C Provenance Working Group. Provenance Working Group Wiki Main Page [EB/OL]. [2012-06-21].http://www.w3.org/2011/prov/wiki/Main_Page.
[16] Beedham H, Missen J, Palmer M,et al. Assessment Of Ukda And Tna Compliance With OAIS And Mets Standards [EB/OL].
[2015-03-15].http://www.webarchive.org.uk/wayback/archive/20140615012529/http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf.
[17]Provenance management[EB/OL].
[2015-03-15].http://www.taverna.org.uk/documentation/taverna-2-x/provenance/
[18]Missier P, Ludäscher B, Dey S, et al. Golden trail: Retrieving the data history that matters from a comprehensive provenance repository[J]. International Journal of Digital Curation, 2012, 7(1): 139-150.
[19] METS Profiles[EB/OL]. [2015-03-15].http://www.loc.gov/standards/mets/mets-profiles.html.
[20] Standard D R, Book R. XML Formatted Data Unit (XFDU) Structure and Construction Rules[J]. 2006.
[21] Dunckley M, Ronen S, Henis E A, et al. Using XFDU for CASPAR information packaging[J]. OCLC Systems & Services: International digital library perspectives, 2010, 26(2): 80-93.
[22]ExLibris[EB/OL]. [2015-03-15]. http://www.exlibrisgroup.com/offices.htm

吴振新 女，硕士，中国科学院文献情报中心硕士生导师，研究馆员。 北京 100190
李文燕 女，硕士研究生，中国科学院文献情报中心。E-mail:lwy9831@163.com 北京 100190