长期保存系统监控服务内容框架研究

吴振新 付鸿鹄 马海收 王玉菊

[摘 要] 在对数字资源长期保存生命周期及可信赖认证标准规范进行研究的基础上，对数字资源长期保存过程中的数据变化影响因素进行分析并将之归纳为数据处理过程的影响因素、数据保存过程的影响因素、数据保存基础设施的影响因素和其他非系统因素等几个方面。在此基础上，提出长期保存系统监控服务的内容框架。该框架从监控系统原有静态参数的系统环境监控，监控系统运行时各个环境状况的系统运行监控，保证数字对象生命周期内的真实性、完整性和可理解性的系统内部数字对象监控三个方面对系统监控服务的内容进行深入的分析和探讨。

[关键词] 长期保存 可信赖仓储 监控服务 数字对象
[分类号] G352 G359
DOI: 10.13266/j.issn.0252-3116.2014.03.008

1 引 言

数字资源同社会发展紧密相关，是人类宝贵的文化遗产，需要长期保存和利用。而数字资源的易逝性、对信息技术的严重依赖等特征导致数字资源易于改变和消失，保存系统如何保障存档数字资源的长期可用，始终是长期保存领域研究和探索的重要问题。

为保证数字资源在经过长时间保存后仍然具有真实性、完整性和可理解性，保存系统需要在数字资源的长期保存生命周期内对其进行严格管理，监控数字对象的重要属性不发生改变或在长期保存政策允许的范围内改变，同时也需要对影响其改变的长期保存系统的运行环境和状况进行监控，以构建数字对象长期可信赖的保存环境。

本文通过对数字资源长期保存生命周期以及可信赖认证标准规范的研究，深入分析影响数字资源发生改变的各种因素，初步归纳形成长期保存系统监控服务内容框架，以期为长期保存系统的进一步实践提供有益参考。

2 相关研究现状和进展

2.1 长期保存相关参考模型

2.1.1 OAIS 参考模型 OAIS 参考模型[1]于 1999 年首次发布，旨在为基于长期保存目的的信息系统建立一个参考模型和基本概念框架，以维护信息系统中数字信息的长期保护和可存取，2003 年成为 ISO 的正式标准，目前已经成为数字资源保存领域普遍遵从的标准规范。

该模型规定长期保存系统承担保存信息并确保目标团体可以获取、利用这些信息的责任。这就要求：首先，系统要对保存的信息有足够的控制，确保信息的长期保存；其次，保存系统应遵循保存的政策和流程，确保信息在一切可能发生的偶然事件中得以保存，确保传播的信息是原始版本的授权副本（authenticated copies），或者可以追溯到原件；再次，确保指定的目标团体能够获得保存的信息；最后，确保指定的目标团体能够理解保存的信息。这些要求可以总结为：长期保存系统需要保证数字资源的真实性、完整性和可理解性。

OAIS 参考模型框架（见图 1），将信息包分为 3 种类型，分别为提交信息包（SIP）、存档信息包（AIP）和分发信息包（DIP），这 3 类信息包分别对应数字对象在长期保存生命周期中的不同阶段。OAIS 要求：保存系统确保在摄入过程中正确地接收和存档 SIP；在长期保存过程中，AIP 没有遭受损坏；在分发访问过程中，能够通过 DIP 真实呈现数字对象。

2.1.2 数字资源保存生命周期模型 英国数字资源保管中心（Digital Curation Center，DCC）于 2008 年发布了数字资源保存生命周期模型[2]（见图 2）。该模型展示了数字对象产生、管理、保存的整个生命周期的各个

[作者简介] 吴振新，中国科学院国家科学图书馆研究馆员，硕士生导师，E-mail: wuzx@mail.las.ac.cn；付鸿鹄，中国科学院国家科学图书馆馆员；马海收，亿赞普（北京）科技有限公司工程师；王玉菊，中国科学院国家科学图书馆馆员。
收稿日期：2014-01-07 修回日期：2014-01-20 本文起止页码：51-57，94 本文责任编辑：王善军

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

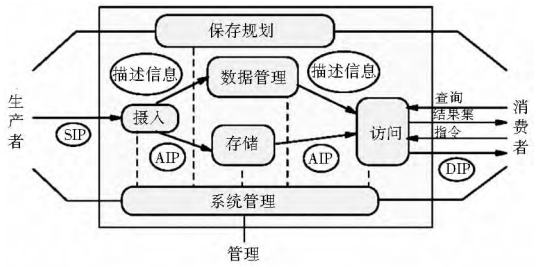

图 1　OAIS 参考模型框架

环节。以数据为核心，该模型从内到外描述了数字资源保存管理的重要方面，分别为描述信息和呈现信息、保存规划、相关群体的关注和参与、存储和长期保存。模型最外层为数字资源生命周期经历的一系列的行为，包括最初的概念化、创建和接收数据、评估和选择、摄入、保存、存储、访问利用和重用以及转换。强调通过对整个保存生命周期的管理来维护数字对象的可信赖性、完整性、可用性。

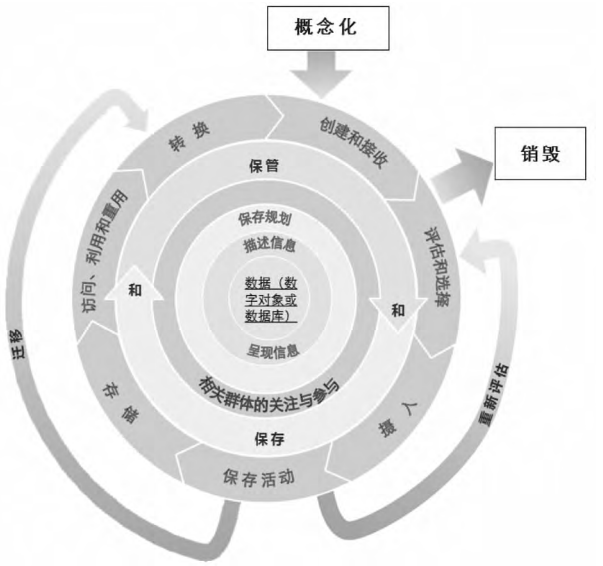

图 2　数字资源保存生命周期模型[2]

该生命周期模型详尽描述数字资源从产生到销毁所经历的各个阶段，为数字资源的保存管理提供了重要的参考依据。模型自发布以来，获得保管（curation）领域广泛的认可和采用，也得到保存（preservation）领域的普遍关注。除了 DCC 采用该模型支持相关的活动，如 SCARP 项目[3]，很多其他相关机构和项目也积极采用该模型，如英国的 UKRDS 项目[4] 和欧盟的 SCAPE 项目[5] 等。

2.1.3　PREMIS 保存元数据数据字典　PREMIS[6] 是支持数字对象长期保存并确保长期可用性的元数据国际标准，由国际专家组成的团队对其持续进行开发维护。该标准在世界各地的数字资源保存项目中得到广泛实施，同时也被融入了许多商业和开放源代码的数字保存工具和系统。

PREMIS 保存元数据字典在 OAIS 信息模型的基础上制定了一个扩展的概念结构，并定义了一个元数据元素集合，此集合可以映射到 OAIS 模型中的概念结构并反映其中的信息概念和需求，但比 OAIS 提供的信息类型更为具体，且使用不同的术语体系。PREMIS 工作组提出了在数字保存活动中涉及到的 5 种实体类型：数字对象、知识实体、事件、行为者和权利。实体包含两方面的信息：一是与其他实体之间的关系，二是实体属性，是实体本身所包含的特征。PREMIS 对可能影响数字对象改变的事件、行为提供了详细的信息记录，从另一个角度反映了长期保存对于数字对象本身以及对数字对象所实施的（可能改变数字对象）操作的描述和记录要求。

2.2　长期保存可信赖认证的要求

长期保存活动中的所有行为可信赖，是数字对象经过长期保存后依旧为目标团体可信赖的基础和保障，长期保存研究的初期就开始了可信赖的研究，并逐步形成了一系列的研究成果。

2002 年 RLG 和 OCLC 联合发布了《可信赖的数字仓储：属性和责任》[7]，2007 年由 OCLC 正式发布了《可信赖数字仓储审核与认证：指标体系与核查表》（TRAC）[8]，经过广泛合作和实验性审计验证，最终在 2012 年发布成为 ISO 正式标准《可信赖数字仓储审核与认证：指标体系与核查表》ISO 16363[9]。

2006 年德国 Nestor 项目可信赖数字保存系统验证研究工作组发表《可信赖数字仓储指标目录》[10]。经过广泛合作和实验性审计验证，在 2011 年正式发布为德国国家标准《可信赖数字仓储指标目录》DIN 31644[11]。

这两个可信赖认证标准为长期保存活动提供了重要的参考和指导。由于 DIN 31644 只有德文版，所以本文采用 Nestor 认证指标目录与 ISO 16363 进行目录层面的对比（见表 1），使大家对这两个标准有个概要性的了解。

2.3　现有长期保存系统的监控功能分析

为了了解在保存实践中监测的应用情况，笔者对目前国际上比较著名的几个保存系统进行了分析。

斯坦福大学的 LOCKSS 采用由多个结点组成的对等网络系统（peer-to-peer），其中的每一个结点称为镜像（box），使用了 opinion polls 这样一种机制，利用存档相同内容的多个结点定期进行内容文档的比较和监控，

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1 Nestor 与 ISO 16363 的认证指标目录比较

	ISO 16363 认证指标目录		Nestor 认证指标目录	
指标体系	1. 组织的基础设施	管理和组织活力 组织结构和人员 过程的可说明性及保存政策框架 财务的可持续性 合同,许可及责任	A 组织框架	系统明确定义了目标 仓储系统允许其指定用户对数字对象所呈现的信息进行适度使用 遵守法律与合同规定 组织形式能够充分满足仓储系统要求 实施高效的质量管理
	2. 数字对象管理	摄入: 获取内容 摄入生产的 AIP 保存规划 信息管理 访问管理	B 数字对象管理	仓储系统能够保证所有处理过程中数字对象的完整性 仓储系统能够保证所有处理过程中数字对象的真实性 仓储系统对其长期保存技术策略有战略规划 仓储系统遵循规定的准则对生产商的数字对象进行转换 依据明确定义的准则,执行数字对象的存档 依据明确定义的准则,进行仓储系统使用 数据管理应满足仓储系统必要功能的实现
	3. 基础设施和安全风险管理	技术基础设施风险管理 安全风险管理	C 基础设施和安全性	IT 基础设施是完备的 基础设施能够保障仓储系统及其数字对象的安全

采用约定的算法来计算各自指定存档内容的信息摘要并统计投票(polling) ,以此来判断存档内容的正确与否。当一个结点发现其内容与其他结点的内容不一致时,该结点会发出请求并选择内容正确的一个结点实现修复。这种投票策略为保证数字内容的不变性和完整性提供了借鉴[12] 。

Fedora 即灵活可扩展的数字对象仓储架构,是康奈尔大学研究的对数字内容进行存储、管理及获取的开源仓储系统,目前被许多机构的保存活动所采用。Fedora 采用了内容版本管理功能来记录数字对象的变化,即当数字对象的 Datastream 被用户修改时,系统会为该 Datastream 创建一个新的版本。Fedora 会保存 Datastreams 的所有版本,这样就保存了数字对象如何随时间变化的历史过程。同时 Fedora 提供了一个特殊的数据流——“AUDIT”Datastream,用来记录该数字对象的所有修改操作,该数据不能被编辑,只受系统控制。每个数字对象的 AUDIT 数据流为操作创建审核记录,记录对象变化涉及到的人、时间、修改内容、原因。另外,Fedora 使用 MD5 验证其数字对象的不变性。Fedora 可以为每个 Datastream 的每个版本生成并保存 MD5,以便后期实现数字对象的不变性校验[13] 。

DAITSS 通过存储池(silo pool) 的方式管理存储。每个存储池有一个数据库用来保存本地副本的原始校验码和最新计算的校验码以及一些历史性信息,包括初次存储时间、所有的不变性检查、删除时间、包丢失时间等。DAITSS 存储服务维护多个存储池,包的多个副本保存在不同的存储池中,一个存储池中一个包最多有一个副本。DAITSS 的保存数据库中保存所有存档数据的保存元数据和操作元数据,包括所有存档包的不变性数据(fixity data) 。DAITSS 的不变性检查包括存储端不变性计算(silo-side fixity computation) 和重新评价(reconciliation) 两个子过程,分别由对应的脚本应用执行。存储端不变性计算对存储池中的每个副本计算新的校验码(MD5 和 SHA1) ,更新存储池数据库并报告不一致情况,包括不变性检查失败、丢包、错包等。对于存档的每个包重新评价包括比较每个副本的校验码及 DAITSS 保存库中的校验码是否一致、是否保存了要求的副本数、是否全部丢失或没有副本、更新 DAITSS 保存数据库中不变性事件记录以反映检查结果、生成检查结果报告[14] 。

Scout[15] 是 SCAPE 项目[5] 中开发的一个相对独立的保存监测系统。它提供一个本体知识库来集中管理监测系统中的保存风险和规避风险所必需的所有信息,采用插件方式集成新的信息源,如文件格式登记系统、特征分析工具、迁移和质量保证、政策、人文知识等。系统可以很容易地浏览知识库并安装触发器以自动通知用户新的风险和机遇。如: 通知用户内容不符合定义的策略,格式变得过时或能够使数据内容重新可用的新工具。Scout 能够监控内容配置文件、允许上传保存控制政策、通过 SPARQL 端点监测 PRONOM 注册表的变化。此外,Scout 项目组在 2013 年的 iPRES 会议上提出了利用信息提取进行 Web 上自动保存监测的实验[16] 。

其他还有 Portico、荷兰国家图书馆的 e-Depot 等,但笔者没有找到相关的资料。可以看出保存监控虽然是保存活动中非常重要的一部分工作,但还没有形成成熟的、系统化的研究成果,更没有比较全面的系统层面的实现。

3 数字对象变化之影响因素的主要分类

导致数字对象重要属性发生改变的因素很多,有来自系统外部的技术发展和目标群体的改变,有系统

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

运行的软件环境、硬件环境、网络环境和外部环境的变化，也有系统内部自身对数字资源的直接管理，这些都会导致比特(bit) 损坏或者引起数字对象内容的改变，因此长期保存系统需要监控的因素也就很多。仔细分析相关参考模型和可信赖认证标准及各保存系统的实际应用，数字对象变化的影响因素主要包括以下几类:

3.1 数据处理过程的影响因素

数据处理过程包括数据检查、摄入、分发等多个方面。在数据摄入到保存系统的过程中需要对数据进行检查、校验、转换、规范等一系列的处理。同时还有随着环境的变化，保存系统面临软件故障和软件过时等风险，为保证数据的可用，需要进行数据格式迁移、软件环境迁移等处理。

3.2 数据保存过程的影响因素

数据保存过程主要涉及数据的存储，包括存储的介质、备份的数量和备份的介质等相关方面。该过程中存在介质老化、故障、过时等可能的风险，为保证数据安全，需要适时进行数据备份、介质迁移等。

3.3 数据分发过程的影响因素

在分发访问过程中，为了能够按照用户需要真实呈现数字对象而进行的处理、转换、传输、呈现等操作。

3.4 数据保存基础设施影响因素

数据保存基础设施包括硬件环境、网络环境等方面。保存基础设施硬件故障和过时、网络服务失败和通讯故障等风险，都会影响数据的安全性和可访问性。

3.5 其他非系统因素

包括人为操作失误、系统遭遇到攻击等多个方面。

4 监控服务内容框架

通过上述研究和分析，笔者初步总结出监控服务内容框架(见表 2) 。

4.1 系统环境监控内容

系统环境，是指在长期保存系统运行过程中不会随时间发生动态变化的因素，属于系统的静态特征，也是保存基础设施部分。系统环境包括外部硬件环境、外部软件环境、内部软件环境和网络环境 4 个部分。硬件包括存储介质、中央处理器和内存等，软件包括操作系统、系统软件、杀毒软件、防火墙以及保存系统需要用到的其他软件等，这些因素是保存系统最基础的部分。此外，保存系统还需要网络传输，因此网络设备也是组成系统运行环境的不可缺少的一部分。系统运行环境作为长期保存系统的基础，其地位十分重要。

4.2 系统运行动态监控

系统运行动态部分，是除与数字对象的生命周期直接相关外，随时间和保存系统的运行不断变化的因素，包括服务器运行的外部环境、硬件、外部软件、内部软件和网络环境 5 个部分。在系统运行时，很多因素的变化都会影响到系统的正常运行。而这些因素发生异常，也可以说明系统的运行出现了或即将出现异常。因此，保存系统需要实时监控这些因素，并为这些因素设定一定的警戒线，当达到警戒线时，向管理人员发出通知，及时处理以避免系统向更坏的方向发展。

表 2 监控服务内容框架

一级分类	二级分类	三级分类
系统环境监控: 系统原有的静态参数	外部硬件环境	存储介质
		中央处理器
		内存
		硬件的更换
	外部软件环境	操作系统
		系统软件
		杀毒软件
		防火墙
	内部软件环境	保存系统用到的软件列表，如数据库软件
	网络环境	网络性能指标
		网络设备
系统运行监控: 监控系统在运行时各个环境的状况	外部环境	服务器外部的环境监控
	硬件环境	存储介质
		中央处理器
		内存
		操作系统
	外部软件环境	杀毒软件
		防火墙
	内部软件及服务运行	内部软件运行状况
		内部服务运行状况
	网络环境	网络设备运行状况
		网络质量监控
系统内部数字对象监控: 保证数字对象生命周期内的真实性、完整性和可理解性	数字对象的摄入过程	数字对象的接收过程
		数字对象的摄入过程
	数字对象的存档过程	AIP 包的监测
		AIP 包修改行为监测
		备份监测
	数字对象的访问过程	用户访问行为监测
		DIP 真实性完整性监测

4.2.1 外部环境 指保存系统服务器所处的自然环境，即服务器机房的环境，主要指标包括温度和湿度。

4.2.2 硬件因素 指在系统运行过程中对硬件的运行状况的实时监控。监控部分关键的硬件指标直接反映系统运行状况，有时甚至可以预测系统出现的问题。例如监控存储设备状况，为存储设备设置最小空间警戒线，防止空间不足导致任务失败。

4.2.3 外部软件因素 主要是指对外部软件的各种修改操作，如操作系统、系统软件、杀毒软件和防火墙等，建议系统记录并进行软件使用期限及权限的管理。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.2.4 内部软件因素 内部软件环境包括长期保存系统及其依赖的软件,如数据库等。需要监控保存系统的各种操作,如记录管理员或操作人的日志,保存系统启动和停止服务应记录时间和原因。同时,系统应监控长期保存各个子服务的可用性,同时也需要监控数据库的各种操作。

4.2.5 网络环境因素 主要指系统运行过程中网络设备的运行状况和网络质量的实时监控,因为这关系到保存系统处理网络传输的效率。

4.3 系统内数字对象监控

数字对象的变化和损坏是影响数字对象的真实性、完整性和可理解性的最直接和最主要因素,因此是整个监控内容框架中的核心和重点,对数字对象变化和损坏的监测是目前保存系统中实施的主要监测功能。根据OAIS参考模型和数据处理流程,将数字对象的监控分为摄入、存档和分发3个部分。系统对数字对象的监控除保证其真实可信外,还起到追踪数字处理流程的作用,一旦部分数据出现问题,系统可以帮助管理人员快速找出问题所在并解决问题,将危害和影响降到最低。

4.3.1 摄入过程监控 数字资源的摄入过程主要包括数据接收和数字对象的摄入两个部分。

(1)数据接收。指系统管理员从数据提供商(一般为出版商)处接收数字资源的过程监控。在接收数据前,保存方应已经与数据提供商就数据的提供达成协议,包括系统需要生产商提供的元数据信息。在接收数据后,管理人员应对接收的数据进行相关检测和监控,验证传输过程的正确性。包括:

- 接收行为监控。接收数据后,系统应记录接收行为的相关信息。
- 原始接收数据备份。接收原始数据后,系统应立即对原始数据进行备份。
- 数据预处理。系统需监控预处理的结果和相关信息。
- 统计相关数据。对接收的数据进行统计,与数据提供商提供的数据进行比较。
- 验证数据接收包。验证所接收数据包的不变性和完整性。
- 验证SIP结构。在SIP真实完整的前提下,根据达成的协议,SIP应包含协议中规定的所有信息。
- 验证SIP中的位流文件。首先系统应检验这些文件的真实性、完整性,避免数据传输过程中位流文件发生改变。此外还应验证文件的格式,是否满足协议要求以及根据长期保存的策略是否满足系统要求,对于不满足长期保存要求的格式,应在保存原有文件的基础上做必要的格式转换。

(2)数字对象的摄入。在确认数据提供商提供的数据正确无误且对其进行必要处理后,开始数据的摄入过程。数据摄入是指将SIP包转换为AIP包的过程。在数据的摄入过程中应监控的因素有:

- 生成相关元数据,包括描述元数据、管理元数据、技术元数据、保存元数据等。在摄入过程中,根据系统定义的AIP包的结构,抽取或生成保存系统需要的元数据,并记录过程中的异常情况。
- 监控位流文件。监控摄入中位流文件的完整性和功能性。
- 摄入完成后,要对新生成的AIP进行真实性和完整性验证。
- 摄入数据备份。系统对新摄入且符合系统要求的AIP进行备份,记录AIP的持久标识符、存储路径、存储时间、提交人等信息。
- 系统记录摄入行为的相关信息,如操作人、时间、摄入条目、成功条目、失败条目、备份位置等相关信息,并发送摄入报告。

4.3.2 存档过程监控 数据的存档过程是指数字对象经摄入流程后在长期保存系统内存储的过程。数据存档在数字资源的生命周期中占绝大部分时间,因此对存档过程的管理和监控十分关键。存档的数据在系统中是以AIP的形式存在的,对存档过程的监控分为AIP的监控、AIP动态修改行为监控和备份监控四部分。

(1)AIP的监控。系统应对存储在系统中的AIP包进行检测,监控其是否符合长期保存的保存标准。对存档AIP的监控可以分为可读性监控、真实性监控、完整性监控和位流文件的监控四部分。

- AIP的可读性监控。AIP在指定位置存在且可读是保证AIP安全存储的最基本的要求,也是对AIP进行其他验证的前提。系统应定期对AIP的可读性进行验证,并向管理人员提交监控报告。
- AIP的真实性监控。验证整个AIP和所有元数据的真实性,其中AIP中所包含的各种元数据都要分别验证。目前主流的方法是消息摘要(message digest),又称数字摘要。
- AIP的完整性监控。AIP作为一个整体,其中可能会包括元数据文档、内容文档以及各种附加文档,系统需要保证AIP包整体的结构完整和关系完整。首

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

先，系统要检测 AIP 声明包含的所有位流文件和元数据是否存在，根据生成 AIP 的准则，完成数据包应含项目的检验和确定，以保证 AIP 的结构完整性。其次，系统要检查 AIP 声明的位流文件和元数据之间的关系是存在的、正确的并且是完整的，以确保 AIP 的关系完整性。检测出的任何异常都需要通知管理人员，并尽快采取相应措施。

• 位流文件的监控。位流文件是指 AIP 中除元数据之外的文件，如 PDF 文档、图片等。监控位流文件有以下几个方面：根据存储信息查找文件的存储位置是否正确和存储介质是否可用、文件是否可读；在位流文件存在的基础上保证文件的真实性，可以参考上文提到的消息摘要的方法；保证位流文件的完整性和功能性，功能性指文件可以被指定的程序读取文件内容并正确呈现；对位流文件格式进行监控，提取格式的相关信息，如格式名称、版本、版本注册信息等，并统计保存系统中各种格式的文件数量，监控格式主要是因为防止文件格式过时。系统可根据实际情况定义格式标准，将禁止的格式和鼓励使用的格式存储到数据库中。

(2) AIP 动态修改行为监控。数据被摄入到保存系统并存储，并不意味着 AIP 会永远不变。保存系统通常允许对原始文档进行修改，但要保证原始文档是可追踪的，修改的过程是可逆的。因此保存系统需要监控修改 AIP 的行为，并对 AIP 做必要的处理。

• 对修改行为的监控。在发生修改事件时，系统要详细记录事件的各个要素。对事件的记录可以参考 PREMIS 保存元数据的事件实体。事件的属性包括事件标识符、事件类型、事件发生日期、事件细节、事件结果信息等属性。系统记录的事件属性元数据应作为原有 AIP 元数据的一部分进行长期保存。

• 对数字对象的追踪。系统在监控修改事件的同时，应对被修改的数字对象做出相应的处理。系统应在原有数字对象的基础上根据事件做出相应的修改，并将其作为数字对象的新版本与历史版本一起存储在保存系统中，完整地记录数字对象的历史演变过程，保证原始文档的可追踪和修改过程的可逆。

(3) 备份监控。要构建可信赖的长期保存系统，建立相应的备份策略是必备的。前文对 TRAC 可信赖认证标准的论述中已经提到备份机制。备份机制可以分为对接收数据的备份和对存档数据的备份。前文已涉及到接收数据以及新摄入 AIP 的备份，在此只介绍对存档数据所做备份的监控。

• 备份 AIP 全部存在且可读。系统应首先确保 AIP 对应的全部备份是存在且是可读的，应根据数据摄入成功后的数据备份信息，找出 AIP 对应的全部备份信息且对其可读性进行验证。如果其中有备份数据已经丢失或不可读，如缺失 AIP 永久标识符、时间、所属批次、存储路径等信息，应立即补充缺失的备份数据并提交报告。系统还应检验所属同一批次或存储路径相同的备份数据包，验证其是否存在。

• 验证其真实性和完整性。前文在 AIP 监控中已经对验证其真实性和完整性做了介绍，处理备份的策略与其不同，因为一个 AIP 包可能存储 3 个或以上的备份数据，对这些备份数据都进行处理势必会浪费大量时间且消耗大量资源。以消息摘要为例，在确认源 AIP 通过验证且确定没有问题的情况下，计算备份数据的消息摘要并将其与源 AIP 的消息摘要进行对比，如果相同则认为备份数据完好。如果在不能确定源 AIP 是否正确的情况下，可参考 LOCKSS 的投票策略[12]，分别计算每个备份 AIP 的消息摘要，并取得票多者为正确的 AIP。但系统需定义得票个数或百分比，以系统有 4 份备份为例(加上源 AIP 一共 5 份)，定义得票个数为 3，即只有当全部 5 份 AIP 中有一种得票大于等于 3，才认定这个数据包为正确数据包，同理可定义百分比如 60%，也即只有在所有 AIP 中得票超过 60% 才能认定为正确。

• 保持 AIP 与其备份的一致性。当发生 AIP 动态修改行为时，AIP 的备份文件也应该与源 AIP 一起发生改变，保持数据的一致性。在修改前，系统应验证 AIP 及其备份文件的一致性，如不一致可采取上文投票策略。在保证其一致性后，再对 AIP 及其备份进行修改，修改完成后再进行一致性验证。当修改前检测出不一致时，系统应向管理员发送错误报告，并对所属批次以及路径相同的备份进行验证，后续验证可由管理员手动执行，也可以采取相应策略由系统自动执行。执行后发送结果报告。

4.3.3 访问过程监控 长期保存系统的建立初衷是保存数据，但其最终目的是保证所保存的信息能够为用户所用，而访问过程就是用户获取和使用数据，是保存系统和数字资源生命周期中很重要的一环。系统对访问过程的监控，需要从监控访问行为和监控 DIP 的真实性完整性两个方面予以保证。

(1) 监控访问行为。访问行为是指用户访问保存系统获取需要的信息的过程。该过程包括对访问行为的安全性和访问功能是否正常的监控，以及对访问日

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

志的记录分析。

• 访问行为记录。系统应记录用户访问行为的信息，如用户名、时间、进行的操作、获取的信息、是否成功、IP地址等。记录这些信息的目的是便于分析统计用户访问行为，并为访问的其他监控行为提供依据。

• 监控安全异常情况。系统应监控访问行为中的异常情况，如未授权的用户或IP地址对数据进行操作或试图更改系统配置，用户进行非法操作，这涉及到系统信息的安全管理。如发生安全异常情况，系统除记录异常日志外，还应向管理人员发出警告。系统可定义安全异常情况连续发生的次数限制，因为一次异常可能是因为用户的误操作，如发生多次则说明是故意破坏，应警告管理人员做出相应处理。

• 监控响应异常情况。系统应监控每次用户请求的响应结果和响应时间。如果出现大范围的获取失败或时间延迟，应立即通知相关人员。系统可以根据经验灵活设置单位时间内的获取失败次数和时间延迟的合理区间。响应异常情况可能由多种原因引起，发生后应立即通知相关人员，及时做出相应处理。

(2) 监控DIP的真实性和完整性，保证用户获得的数据真实可信。

• 完整性监控。此时的完整性与存档过程中的完整性意义不同，DIP是根据用户的请求并根据AIP生成的，因此访问过程的完整性监控应根据用户的请求清单对DIP的内容进行完整性检测，确保DIP包含用户请求的全部内容。

• 真实性监控。在包含用户请求的全部内容的基础上，系统应对DIP中包含的每种元数据和文件与AIP中的元数据和文件做真实性校验，确保DIP全部元素的真实性。

5 结 语

影响数字对象重要属性发生改变的因素有很多，在数字对象保存的完整生命周期内，如何监控和记录这些因素对长期保存环境和数字对象产生的影响，对于保证数字对象的真实性、完整性和可理解性至关重要。本文主要从系统运行的环境以及系统内部的存档管理方面进行考虑，对该领域的研究现状及进展进行了研究和分析，基于长期保存相关参考模型和可信赖认证标准，初步总结了一个比较完整的长期保存系统监控服务内容框架。由于篇幅的限制，实现监控服务所采用的具体技术和方法以及相关长期保存系统的实践应用情况，本文没有涉及。

总体来讲，保存系统和保存对象的可信赖是保存领域的核心目标，监控服务能够在一定程度上为此提供保障，这个领域未来还有很多工作需要深入探索，希望本文能为长期保存系统在保存生命周期中维护和管理存档对象提供一定的借鉴和参考。

参考文献：

[1] Reference model for an open archival information system [OL]. [2014-01-20]. http://public.ccsds.org/publications/archive/650x0m2.pdf.

[2] DCC Curation Lifecycle Model [OL]. [2014-01-20]. http://www.dcc.ac.uk/sites/default/files/documents/publications/DCC-CLifecycle.pdf.

[3] SCARP [OL]. [2014-01-20]. http://www.dcc.ac.uk/projects/scarp.

[4] UKRDS [OL]. [2014-01-20]. http://www.jisc.ac.uk/news/managing-uk-research-data-for-future-use-05-mar-2009.

[5] SCAPE [OL]. [2014-01-20]. http://www.scape-project.eu/.

[6] PREMIS data dictionary for preservation metadata [OL]. [2014-01-20]. http://www.loc.gov/standards/premis/v2/premis-2-1.pdf.

[7] RLG/OCLC trusted digital repositories: Attributes and responsibilities [OL]. [2014-01-20]. http://www.prestocentre.org/system/files/library/resource/tdr%20attributes%20and%20responsibilities.pdf.

[8] Trustworthy repositories audit & certification: Criteria and checklist [OL]. [2014-01-20]. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

[9] Space data and information transfer systems — Audit and certification of trustworthy digital repositories [OL]. [2014-01-20]. http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510.

[10] Catalogue of criteria for trusted digital repositories [OL]. [2014-01-20]. http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf.

[11] DIN 31644 [OL]. [2014-01-20]. http://www.nabd.din.de/cmd?level=tpl-art-detailansicht&committeeid=54738855&artid=147058907&languageid=de&bcrumblevel=4&subcommitteeid=112656173.

[12] LOCKSS [OL]. [2014-01-20]. http://static.usenix.org/events/usenix2000/freenix/full_papers/rosenthal/rosenthal.pdf.

[13] Fedora [OL]. [2014-01-20]. http://www.fedora-commons.org/.

[14] DAITSS [OL]. [2014-01-20]. http://daitss.fcla.edu/.

[15] Scout [OL]. [2014-01-20]. http://www.openplanetsfoundation.org/blogs/2013-12-16-scout-preservation-watch-system.

[16] Automatic preservation watch using information extraction on the Web [OL]. [2014-01-20]. http://purl.pt/24107/1/iPres2013_PDF/Automatic Preservation Watch using Information Extraction on the Web.pdf.

(下转第94页)

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

录情况、检索系统性能及功能等方面，但是，由于数据库的差异，比较的项目和内容在不同数据库间差别很大，因此有时也难以做到客观和准确的评估。

本文针对西南大学综合性研究型大学的特点，选择比较了4个综合性的西文全文期刊数据库，它们所涵盖的学科范围都比较广泛，对科学技术、医学、社会科学及人文科学方面的期刊都有所收录。探索性地对各综合性数据库所含信息量进行横向比较，仅从数据库的客观事实特征这一方面为图书馆对数据库的了解提供了一个视角。而信息量与用户和性价比关系的分析，也仅是帮助图书馆了解用户对数据库的满意度，而要全面、客观且真实地评估电子资源，还需要将数据库与西南大学的教育课程和科研的相关程度相比较，进行用户满意度调查，将用户对数据库的使用效率、利用效果与用户真实需求一起联系进行评价。

参考文献：

[1] 史慧丹，周蕊，邬淑珍. 我国电子期刊全文数据库评价研究文献学综述[J]. 现代情报，2007，27(9)：15－17.

[2] 周荫清. 信息理论基础[M]. 北京：北京航空航天大学出版社，2002：167.

[3] 上海市科学技术编译馆. 信息论理论基础[M]. 上海：上海市科学技术编译馆，1965：1－35.

[4] 维基百科. 香农多样性指数[EB/OL]. [2013－12－15]. http://baike.baidu.com/view/3077672.htm.

[5] 维基百科. 多样性指数[EB/OL]. http://zh.wikipedia.org/zh-tw/%E5%A4%9A%E6%A0%B7%E6%80%A7%E6%8C%87%E6%95%B0.

[6] Pielou B C. Ecological diversity[M]. New York：John Wiley & Sons，1975.

[7] 王轶帅，陆思霖. 国外综合性网络全文数据库的特点及其对图书馆的启示[J]. 科技情报开发与经济，2009，19(30)：5－7.

[8] 黄永礼. 电子文献数据库的评估与选购[J]. 大学图书情报学刊，2003(2)：50－52.

[9] 胡乃志. 高校图书馆数据库评价与选择的策略研究[D]. 长春：东北师范大学，2007.

The Research of Foreign Journals Full-text Database Evaluation Based on Information Diversity Index

Yang Fan　Li Xuemei

Library of Southwestern University，Chongqing 400715

[Abstract] With four periodicals full-text database Elsevier，Springer，EBSCOhost and JohnWiley，for example，using the richness index (S)，Shannon diversity index (H) and the Pielou evenness index (J) to evaluate and compare the information amount of several databases. Data analysis showed that the S and S_1 of ASP publication of EBSCOhost are the highest in several databases and H and J is the lowest. Analysis the relationship between information index and the user usage，and Analysis shows that the Elsevier database H and J from the highest value，the user visits and full-text downloads is the highest，and most highest performance-to-price ratio. The research results can provide perspective for library knowledge database.

[Keywords] The amount of information　Diversity index　Database　Assessment

(上接第57页)

Study on Monitoring Service Frameworks of Digital Preservation Systems

Wu Zhenxin[1]　Fu Honghu[1]　Ma Haishou[2]　Wang Yuju[1]

[1]National Science Library，Chinese Academy of Sciences，Beijing 100190

[2]IZP Technologies Co.，Ltd.，Beijing 100102

[Abstract] Based on research of digital preservation life cycle and trusted audits and certification standards，the authors gave an in-depth analysis of a variety of factors which may make key attributes of digital object change. The factors can be divided into several categories such as data processing，data storage，infrastructure and some other aspects. On this basis，the authors summarized the monitoring service framework of digital preservation systems which include environmental monitoring，run-time system performance monitoring and digital objects monitoring.

[Keywords] digital preservation　trustworthy repository　monitoring service　digital object

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net