2011 年第 8 期（总第 8 期）

长期保存跟踪扫描

主办单位：中国科学院国家科学图书馆

2011 年 10 月

为传播科学知识，促进业界交流，特编译《长期保存跟踪扫描》，仅供个人学习、研究使用。

目　录

【专题报道】 1
欧盟 ICT 计划共同资助数字资源长期保存项目研究报告 1
【重要文献摘译】 15
《分布式数字资源长期保存指南》摘译之六 15
【动态追踪】 21
JISC 学术研究完整性会议 21
翻译的数字化实践：OAAP 项目的启示 21
SDSC 推出可升级、高性能的云数据存储服务 22
FDA 存储库数据量超过 100TB 23
【信息扫描】 24
POCOS 格拉斯哥座谈会－－软件艺术 24
美国国会图书馆开展数字保存拓广与教育计划 24
JISC 服务教学、科研的云计算——已发布的项目和合作伙伴 25
VIDaaS 项目启动 25
21 世纪长期保存路线图：迈向数字未来 26
第四届 eSciDoc Days 会议召开 26

【专题报道】

欧盟 ICT 计划共同资助数字资源长期保存项目研究报告

Stephan Strodl，Petar Petrov，Andreas Rauber 著 么媛媛 编译

摘要：该报告对欧盟委员会基于研究与技术发展的第六、七框架计划赞助的数字资源长期保存研究项目进行了综述，总结了这些项目的目标、发展态势、相似点和不同点以及项目成果。报告还考虑到了数字保存领域的研究议程，并提出了未来所面临的挑战。

1 引言

在刚刚过去的 30 年中，信息技术极大地改变了我们思考、创造、存储、表示和共享信息的方法。这一迅速且剧烈的变化带来了许多无法想象的进步、发现和创造，但也给我们带来了长期保存数字资源的重大难题，不管是新获取的知识还是我们的文化遗产。

1.1 数字资源长期保存的难题

30 年前几乎所有资料都是纸质的，只要条件合适，人们就能长期保存它们。从长期存储的角度而言，如今的存储介质不如纸质媒介可靠。技术飞速革新促进了更快的硬件、更高效健壮的软件和大量的新格式的发明。因此以旧软件执行、旧格式编码、存储于旧媒介上的内容很难有机会保存超过十年。

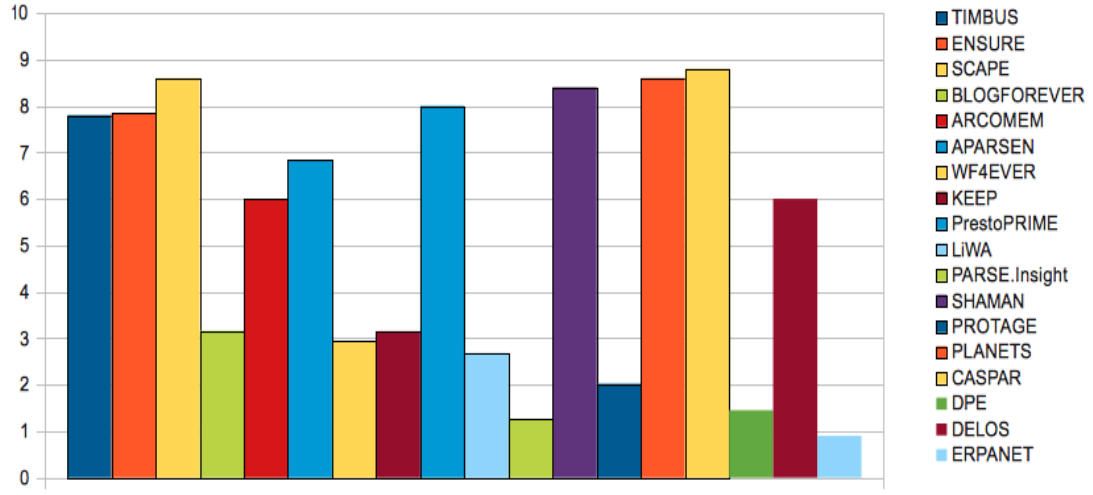
幸运的是，即使工商业活动中常常被忽略，欧盟委员会在早期就发现了这一保存问题。为了防止数据流失，欧盟基于研究和技术发展的框架计划资助了部分数字资源长期保存的研究项目。

1.2 项目综述

1.2.1 重要图表

在欧盟的第六和第七框架计划中共有超过 9000 万欧元被投入到数字资源长期保存相关课题研究中，惠及 15 个关注数字资源长期保存的项目，其研究伙伴来自超过 20 个国家。FP6（第六框架计划）始于 2002 年终于 2006 年，FP7（第七框架计划）预计从 2007 年进行到 2013 年。

图表 1 做了一个项目投资概述。需要注意的是，第七框架计划中的数字资源长期保存项目投资至少扩大至 3 倍，这表明其对这一问题的重视。随着财政支持力度的加大，不仅研究项目的数量增加了，而且倡导数字资源长期保存的权益人和机构也变多了，促进了多个领域和机构间建立坚实的数字资源长期保存（DP）专业体系。



图表 1：欧盟对所有数字化保存项目的投资　　单位：欧元

1.2.2 ICT 计划中的数字资源长期保存项目

以下的项目构成了第二部分进行评估的基础，项目简短介绍大多摘自项目网站，项目详细描述和重要特点见附录。

现行项目（按日期倒序排列）

- **TIMBUS（FP7，IP）**该项目将努力扩大 DP 的外延，包括一整套能确保持续访问服务和软件的活动、流程和工具，这些服务和软件为信息可以被访问，被正确地表示、验证并转化成知识提供了不可或缺的基础。
- **WF4Ever（FP7，STREP）**该项目的目标是为科学工作流的长期保存提供所需的方法和工具。
- **ENSURE(FP7，IP)**该项目将确保人们能够长期利用数量急剧增长的、商业机构产出或控制的数据。这将通过分析如医疗保健、临床研究、金融服务等不同领域的案例，显著地延伸数字保存技术的应用态势，最新的焦点在相对同质的文化遗产数据上。
- **SCAPE(FP7，IP)**该项目将通过三种方式加强数字资源长期保存的技术水平：为可扩展的长期保存活动开发基础设施和工具；为自动化的、保质保量的长期保存工作流提供一个框架；通过政策性长期保存计划和监督体系把这些组件整合在一起。
- **BlogForever（STREP）**该项目为博客资源长期保存、管理和分发创造了一些工具。
- **ARCOMEM (FP7，IP)**该项目的愿景是利用人群的智慧，对内容进行评价、选择和长期保存，使存档反映集体的记忆和社会内容。
- **APARSEN(FP7，NoE)**即欧洲科学记录的联合长期访问（Alliance Permanent Access to the Records of Science in Europe）是一个汇集数字资源长期保存从业人员和研究者的杰出网络。
- **KEEP(FP7，IP)**该项目正在开发仿真服务以确保正确地渲染静态和动态的数字对象——文字、声音和图像，多媒体文件，网站，数据库，电子游戏等。该技术正对早期的电

脑游戏进行测试。

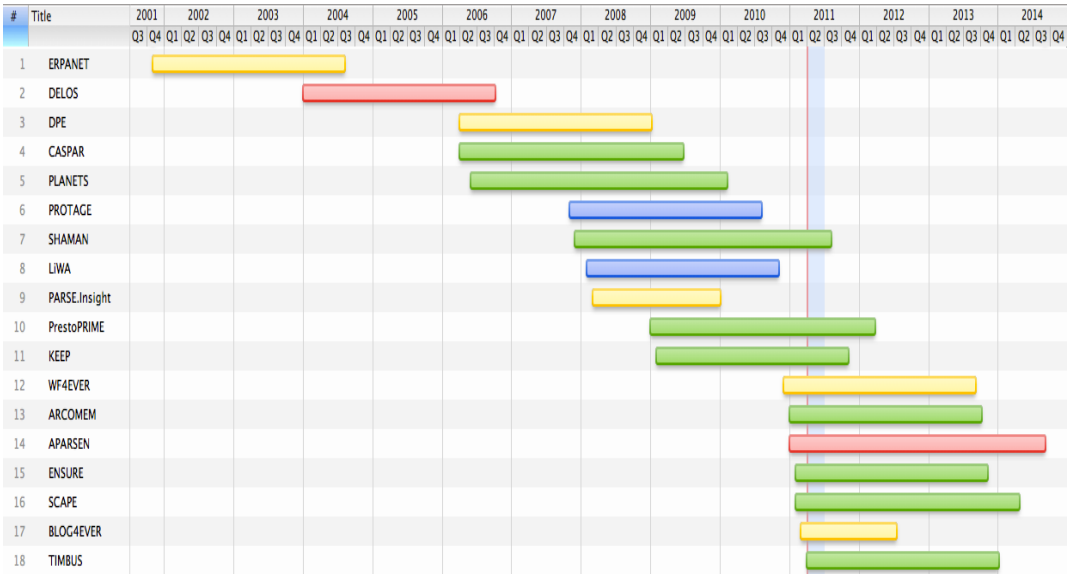
- **PrestoPRIME (FP7，IP)**该项目正与欧洲网络数字图书馆合作通过整合其媒体馆藏来解决数字视听内容的长期保存和获取问题。研究的成果将是通过联网的 Competence Centre PrestoCentre 来发布的一系列工具和服务。
- **LiWA (FP7，STREP)**该项目开发并展示一系列网页存储工具，这些工具能捕获多种来源的内容，能提高存储的精确度和真实性，并确保网络内容的长期可解析性。
- **SHAMAN (FP7，IP)**该项目正在开发新一代的数字资源长期保存框架，包括分析、摄取、管理、访问、再利用信息资源和图书馆、档案馆数据的工具。

过去的项目

- **PARSE.Insight (FP7，CA)**该项目旨在强调数字化研究数据的长存性和多变性，并把研究重点集中在支撑欧盟研究的数字资产的持久性和可解析性所需要的 e-Science 基础设施部分。
- **PROTAGE (FP7，STREP)**即保存机构在代理环境下使用的工具（Preservation Organizations Using Tools in Agent Environments），该项目涉及保存的资源数量不断增多、资源异质性不断增加所带来的挑战，并通过开发能使长期保存流程更加高效和更加有自发性的工具来解决这一问题。
- **PLANETS (FP6，IP)**
- 即我们的文化和科学遗产的长期保存和获取（Preservation and Long-term Access to our Cultural and Scientific Heritage），它的基本目标是构建实践性的服务和工具来确保数字文化和科学资产的长期获取。该项目的成果是为数字信息保存与管理提供了一个集成的生产环境，尤其侧重于图书馆和档案馆的需要。
- **CASPAR (FP6，IP)**即文化、艺术和科学知识的保存、访问和检索（Cultural，Artistic and Scientific Knowledge Preservation，for Access and Retrieval）。CASPAR 团队创造了一个工具和基础设施组件的结构来支撑所有类型的数字编码信息端到端的长期保存，从而帮助生产商、保存人员和用户来分担数字资源长期保存的负担。
- **DPE - DigitalPreservationEurope (FP6，CA)**欧洲数字保存是一个协作项目，推出的目的是提高当前活动中数字资料长期保存的合作性和一致性。该项目提升了数字资源长期保存的地位，提高了数字资源长期保存过程的审核和认证标准，并通过培训促进了长期保存技能的发展。
- **DELOS (FP6，NoE)**即 Network of Excellence on Digital Libraries。该项目的研究领域覆盖了馆藏结构、信息获取和个性化，视听对象和非传统对象，用户界面，知识抽取，语义互操作性，长期保存和评估。
- **ERPANET (FP5)**旨在建立一个可扩展和自我维护的欧洲启动计划，该计划作为一个虚拟的信息交流中心和知识库服务于文化遗产和科学数字对象的保存领域。

1.2.3 项目间联系

数字资源长期保存领域中存在一些项目间的关系。图表 2 是这些项目的一个甘特图（Gantt Chart），蓝色代表“有具体针对性的研究项目”，红色是“优秀网络”，黄色是“协调行动”，绿色是“综合项目”。



图表 2：ICT 计划的时间轴

DELOS 数字资源长期保存集群（DPC）构建于早期 ERPANET 成功工作的基础之上，并继承和发展了 ERPANET 的工作，项目企业集团的主要执行者保持不变。DigitalPreservationEurope(DPE)指的是 ERPANET 和 DELOS DPC 的工作，该项目在概述指导方针和重点课题上继续并延伸了它们的工作。DELOS 项目的评价体系为促成 PLANETS 计划打下了基础，这项工作将在 SCAPE 项目内进行进一步探讨。

最近开始的 APARSEN NoE 项目借鉴了 PARSE-Insight 项目的一些工作如路径图、调查结果，同时也引用了 CASPAR 的一些工作成果。

另一个从音像方面成功地运用并继续发展以前项目成果的例子来自 PRESTO。PrestoSpace 延伸了 PRESTO 广播装置以外的获取方法。这两个项目都开发了保存数字化视听馆藏的方式和方法。最新的项目 PrestoPRIME 建立在 PRESTO 和 PrestoSpace 工作的基础上，侧重于视听内容的数字资源长期保存。PrestoPRIME 解决了“原生数字化”视听内容的保存，在 2011 年 3 月建立了联网的能力中心 PrestoCentre。PrestoCentre 是一个成员驱动型组织，汇集了音像数字化和数字资源长期保存方面的相关利益团体。同样地，Open Planets Foundation (OPF)作为 PLANETS 成员驱动型项目的继任者，为该项目开发的解决方案注入了生命力。

2 研究项目

本节中讨论的项目将根据不同的标准进行分类——如所讨论内容的类型、核心目标、目标受众，以更好地概括它们之间的共性并为研究它们间的关系提供一个新的视角。

2.1 内容类型

从所讨论的内容类型角度来看，这些项目被分为八个主要类型：

Office 文档（包括所有种类的文本文件和图像）、音频/视频内容、科学数据、网络内容、社会网络、互动内容、应用程序和进程。因为 CA SPE 项目没有一个特定的重点研究的内容类型，所以并没有被分类。

Project/ ContentType	Starting Year	Office Documents	Audio/Visual	Scientific data	Web Content	Social Web	Interactive Content	Applications	Processes
TIMBUS	2011							x	x
ENSURE	2011			x					
SCAPE	2011	x		x	x				
BLOGFOREVER	2011				x	x	x		
ARCOMEM	2011				x	x			
APARSEN	2011			x					
WF4EVER	2010			x					x
KEEP	2009						x	x	
PrestoPRIME	2009		x						
LiWA	2008				x				
PARSE.Insight	2008	x		x					
SHAMAN	2007	x	x	x					x
PROTAGE	2007	x							
PLANETS	2006	x	x						
CASPAR	2006	x	x	x	x			x	
DPE	2006								
DELOS	2004	x							
ERPANET	2001	x		x	x				

图表 3：项目的重点研究内容种类

图表 3 中的项目是按它们的启动年份排序的。表中清楚地分析了各个项目研究内容类型的转变，FP5 和 FP6 中的项目研究重点集中在 office 文档上，包括机构设置中的一些图像。这些项目在内容上采取了广泛应用的方法，目的是提供可在多种环境中使用的基本概念和工具。有一些视听资料和科学数据的长期保存工作已经被 PLANTS，CASPAR 和 SHAMAN 等项目完成。SHAMAN 是第一个明确地解释了产品生命周期管理（PLM）的要求和设计与工程领域长期保存工作相关流程的项目。

然而，真正转向研究更复杂的数据结构和格式的项目出现在第七个框架计划（FP7）中。像 LiWA，SCAPE，BLOGFOREVER 和 ARCOMEN 这样研究科学、社会和网络相关内容的项目处理了保存通用文件这一难题的不同方面。ENSURE 项目研究了医疗保健和金融部门的数据处理问题。

从图表 3 中可以看出一个发展趋势：内容类型从左到右变得更为复杂。SCAPE 和 ENSURE 项目是例外，它们处理大型办公文档（SCAPE）和临床试验（ENSURE）的长期保存。

目前的研究重点是互动对象、嵌入对象、本体和临时数据。这种转变在刚启动的 F7IPs TIMBUS 和 WF4EVER 项目分析数字材料时更加明显。TIMBUS 项目研究支持长期保存的

业务流程和应用。WF4EVER 调查实现科学工作流程的长期保存、高效检索和再利用的科技基础设施。这两个项目显示了脱离实体文件后，查看文件数据、结构和性能的范式上的转变。

2.2 目标群体

Project/ Target Institutions	Starting Year	Memory Institutions	Scientific Institutions	Government Organizations	Enterprise	Private
TIMBUS	2011		x		x	
ENSURE	2011				x	x
SCAPE	2011	x	x		x	
BLOGFOREVER	2011	x	x		x	x
ARCOMEM	2011	x				
APARSEN	2011	x	x			x
WF4EVER	2010		x			
KEEP	2009	x				x
PrestoPRIME	2009	x	x			
LiWA	2008	x				
PARSE.Insight	2008		x			
SHAMAN	2007	x	x	x	x	
PROTAGE	2007				x	
PLANETS	2006	x	x	x		
CASPAR	2006	x	x	x		
DPE	2006	x	x	x		
DELOS	2004	x				
ERPANET	2001	x	x	x		

图表 4：项目的目标受众

图表 4 列出了每个项目的目标受众，为研究重点提供了另一种视角。该表将目标受众分为五种：存储机构、科研机构、政府组织、企业和个人。存储机构不仅指图书馆和档案馆，还包括网络档案馆、数字图书馆和数字知识库。个人这一类别包含了不能被分到其它类别中的所有终端用户。

从图表 4 能看出研究的重点主要在存储机构和科研机构上，这并不奇怪，档案馆和图书馆是首当其冲地面临着数字资源长期保存问题的组织，它们有义务长期保存它们的虚拟馆藏和数字馆藏。因此，存储机构在数字资源长期保存领域占有重要地位。SHAMAN 和 PROTAGE 项目首先进行了企业方面的研究。

有一系列项目还没有确定一个特殊的目标受众群体，因为它们采用普遍使用的基本技术研究了一些一般性的问题。

刚刚启动的研究项目 TIMBUS，ENSURE 和 SCAPE 非常明确地把重点转移到了满足业务部门的需求上。TIMBUS 将侧重于业务流程的长期保存，SCAPE 将为大型半自动化工作流编排和一些异构数据集开发可扩展的服务和业务流程平台。ENSURE 将在考虑经济影响的基础上研究可伸缩的量入为出的长期保存基础设施。

2.3 研究伙伴

图表 5（图略）给出了涉及两个或两个以上 ICT 数字资源长期保存项目的合作伙伴列表。第一列表示这个合作伙伴是遗产机构（ALM），科研机构（SCI）还是行业伙伴（IND）。

名单上有一些大的国家级图书馆，这并不奇怪，在以前的项目中只有两个国家级档案馆（荷兰和瑞士档案馆）参加了两个以上的项目。虽然在过去这些图书馆很活跃，但名单显示只有两个新项目（APARSEN 和 SCAPE）中有图书馆的参与。这种趋势在 2.1 和 2.2 中也有所体现。图表 5 列出了参与数字资源长期保存研究项目的大量科研机构，一些德国和英国的大学从事这方面的研究。参与两个或更多研究项目的研究机构数量的庞大表明它们将在其机构内部继续开展数字资源长期保存工作。

图表 5 列出的所有行业伙伴都把为数字资源长期保存提供解决方案和产品作为一项业务。有两个项目（TIMBUS 和 ENSURE）是由行业伙伴领导的，虽然他们是所在企业集团中至今为止唯一的合作伙伴，但他们仍是拥有问题的代表者而不只是解决方案的提供者。

2.4 核心目标

Project/ Objectives	ERPANET	DELOS	DPE	CASPAR	PLANETS	PROTAGE	SHAMAN	LiWA	PARSE.Insight	PrestoPRIME	KEEP	APARSEN	ARCOMEM	BLOGFOREVER	SCAPE	ENSURE	WF4EVER	TIMBUS
Appraisal and Selection	X							X					X					
Characterization				X	X		X			X					X			
Format identification				X						X								
Metadata				X			X	X		X	X		X					
Preservation Action				X	X					X	X	X			X			
Preservation Planning		X			X					X					X			
Authenticity & Trust			X	X		X		X	X	X						X		X
Access		X		X			X		X	X	X							
System design				X		X	X			X						X		X
Workflows				X	X										X	X	X	X
Tool Development				X	X	X	X			X			X	X			X	
Interoperability				X	X									X				
Scalability							X	X		X					X	X		
Legal			X							X	X	X		X				X
Research Roadmap			X						X						X			
Training	X	X	X	X	X							X						
Coordination			X									X						

图表 6：项目的核心目标

2.4.1 重点课题，更广泛的应用与新方法

数字资源长期保存（DP）研究在欧洲的总体发展情况可以用对核心课题和项目目标的概述来描绘。数字资源长期保存领域早期研究项目从基本概念、体制和方法的定义、设计与讨论着手，这些项目主要靠数字图书馆和档案馆社区来驱动。在众多主题中，占主导地位的有：选取和评价、元数据的定义、唯一标识符、描述工具。第一阶段的项目是 ERPANET 和 DELOS。

第二阶段，可用的方法、工具和模块被纳入到框架结构中。融入数字资源长期保存模块的框架结构，能将工作流的各个组成部分整合进其他系统中去。它促进了 DP 工具的广泛运用，例如 PLANETS 可互操作性框架体系，CASPAR 整合框架体系和运用 SHAMAN 网格技

术的集成保存框架体系。

下一轮项目提出了更加专业化的应用情景和新的解决方法。迄今为止开发出的工具和方法都将重点放在 boread 应用情景上，PROTAGE 项目的代理环境和 ARCOMEM 项目中社会网络的应用都是解决问题的新方法。不论是在以网络存档技术为基础的 LiWA 项目的还是在以仿真技术为基础的 KEEP 项目中，集中化和专业化的主题都被提了出来，这在其他研究项目中都很少被讨论到。

2.4.2 可扩展性

数字资源长期保存现有的工具大多有其特定的功能，并没有被设计为能大规模应用的工具。在实践中，我们面对的是数量庞大的内容，如网络存档或大型机构知识库（档案馆或图书馆等）。这些团体都需要可以扩展的工具和方法来处理大量的对象。SCAPE、ENSURE 和 LiWA 都提出了长期保存解决办法的可扩展性，TIMBUS 和 ENSURE 都在研究云存储的可扩展性和长期保存虚拟技术的应用。

2.4.3 智能的工具和方法

第一个被开发出来的工具需要数字资源长期保存的人际合作和渊博知识。现今的趋势向智能工具和辅助用户、支持决策过程的方法方面发展。通过创造并使用知识基础和创新方法，工具可以升到下一级来支持复杂环境和高容量的异构内容。

2.4.4 概念模型和系统设计

OAIS 参考模型及其术语已被确立为一个共同的数字资源长期保存领域中的概念和模型的基础。已有一系列的研究项目在做概念模型的研究工作了。

CASPAR 概念模型极大地受到了 OAIS 模型的影响，它提供了一个通用的基础设施的概念以支持数字资源长期保存。该模型还包括一个信息模型，详细地讨论了表示信息和长期保存描述信息的概念。

有关如何从系统设计角度解决数字资源长期保存问题的第一个构想是由 SHAMAN 项目实现的。该方法确定了系统的主要特点和要求并促进企业结构框架去解决数字资源长期保存问题。SHAMAN 设计了一个基于 OAIS 参考模型的参考体系结构，提出了反映权益人关注重点的几种观点。该模型提供了一个有助于推导出目标环境中数字资源长期保存体系架构的过程。

PLANETS 项目开发了一个捕获需求的概念模型，将风险与减少风险的行为联系起来并从权益人特殊需求角度来解释它们。

PROTAGE 运用了一个不同的系统设计方法——多代理系统结构体系（Multi Agent System Architecture）。该结构体系由软件代理工具和数字资源长期保存与访问的网络服务组成。长期保存过程通过代理来协调，这些代理会自动定位、选择并运用网络服务来获取信息并作出决策、执行保存任务。网络服务为代理和用户提供特定类型的服务，如访问、用做知识基础、病毒检查、迁移和元数据提取服务。

TIMBUS 将探讨智能企业的参考架构保存功能的风险管理系统。该项目中系统设计方面被阐述得更为详细，如为开发可数字化保存的软件服务和系统提供指导方针。

SCAPE 和 ENSURE 项目阐述了长期保存的工作流。该项目的研究对象是数字对象保存工作流的设计与编排。

2.4.5 真实性、信任与审核

所有数字资源长期保存系统的一个共同要求是数字对象的真实性。这包括确保数字对象的完整性，即保证其信息内容没有被修改。目前在对象真实性领域的尝试是建立对真实性和完整性概念的共同理解和理论基础。关于真实性的技术方面，一些研究项目正开发不同种类对象的解决方法。

做为真实性的一部分，保存数字对象的语义层面甚少被研究。LiWA 项目阐述了语义长期保存的另一部分：语言随时间的变化。为了从长远的角度来解释这一内容，他们一直在开发处理术语演变的自动方法。

信任是长期保存的一个重要方面。DPE 是通过开发自我评估方法（DRAMBORA）来解决这一问题的首个项目。它鼓励各个组织在识别、评估和管理组织内部风险之前建立对他们的对象、活动和资产的全面自我认识。

数字知识库的信任和真实性工作已被 CASPAR 项目解决，该项目对真实性问题和 OAIS 中的 DRM 处理进行了升级，它成功地促成了数字资源长期保存 ISO 认证和认证过程的开发。APARSEN 项目开发了知识库的独立的第三方常用认证方法。

认证法律背景下的长期保存问题非常有趣，许多项目都忽视了这一问题。DPE 在这一领域做了一些初步的工作，KEEP 发表了其关于仿真的法律研究。CASPAR 研究了 DRM 并开发了一个应对不断变化的立法基础和适应法律体系的多样性的工具。TIMBUS 将解决保存业务流程的法律问题。

2.4.6 元数据

数字资源长期保存中的元数据工作耗费了人们大量的精力，一些描述性元数据的标准已被设立，如 MAB，MARC，Dublin Core 等够已被广泛接受和运用。数字保存领域 PREMIS 已经被确立为核心保存元数据标准。

虽然各个团体都非常期待一套长期、稳定的元数据标准，并倡导继续演化并延伸元数据标准和纲要。新标准、建议和实施背后的动力是对不同的设置、目标、对象和方法的更好的支持和更广的覆盖。

几乎所有的研究项目都已做完元数据方面的工作了，CASPAR 项目研究对象的描述性信息和表示信息。PLANETS 正在调查其超前的特性并创造关键数字资源长期保存元数据概念的数据词典。底层概念模型支持的是动态的保存过程而不是静态的特征和事件的记录。

元数据的特定应用研究由 KEEP，LiWA，ARCOMEN 和 PrestoPRIME 来进行，以支持模拟器、网络存档、社会网络和音像内容的保存。

2.4.7 语义技术

CASPAR 项目运用语义 Web 技术来塑造数字资源长期保存的知识模型，知识管理服务中所包含的信息与目标团体的表示信息和知识基础有关。SHAMAN 语境模型提供了一个用本体表示数字对象属性和它们之间联系的独立基础设施。

2.5 训练和教育

几乎所有的研究项目都就它们所做的数字资源长期保存工作提供了培训课程。除了正式培训，还需要一个数字保存的专业计划。

2.6 相关学科

近年来，数字资源长期保存本身已成为了独立的科学学科，而跨学科的数据处理与许多其他的学科和主题都有联系和重叠，所以在本节列出了所选择的其他学科的概述。以下是相关学科名称（概述略）。

- 信息管理、数据管理、知识管理
- 信息检索&数据挖掘
- 语义技术
- 企业结构
- 定向服务和云计算
- 数据库
- 存储技术
- 风险管理
- 治理，最佳实践，遵守和信任
- 立法
- 数字版权管理
- 安全性
- 隐私性

3 研究议程

数字资源长期保存的问题大概产生于 20 年前，当时已经有一些独特的保存措施了，随后人们又提出了几个研究议程来系统地解决这些问题。DPE 研究项目总结了早期的议程和 PARSE.Insight 项目的路线图还有 Dagstuhl 研讨会的成果，所以本节以 DPE 为基础展开介绍。Dagstuhl 研讨会试图建立一个专门的 IT 方面需要综合研究的轮廓。PARSE.Insight 项目为科学数据基础设施提供了一个路线图来识别其缺少的组件，这个项目对三个利益相关领域——科研、出版和数据管理进行了大规模的调查研究，调查结果被输入到研究路线图中。这份报告仅列出了研究主题，省略了只关于组织、工程和发展的问题。

3.1DPE 研究议程

DPE 对所有现有的研究议程进行了透彻的分析来总结什么是必须要做的，并确定在数

字资源长期保存条件下最终构成欧洲研究和发展的基础的问题中哪些是被忽视的。共有 10 个研究领域被提出：

- **恢复**：即使有一些可以完全恢复损坏的媒体文件的数字取证技术，但因为文件类型是未知的，如何转化这些媒体文件就成了很大的难题。因此转化这些对象并揭示其真实内容是数字资源长期保存中的巨大挑战。
- **保存**：为了解决出现的诸如迁移和仿真等技术、方法和策略过时退化问题，一种持续保存旧数据的方法被提出了。然而，这给该领域的研究带来了新一轮的挑战和研发的新课题。
- **管理**：规划、制定、执行、管理和监测数字资源长期保存的组织流程必须成为研究的重点。
- **风险**：数字资源长期保存基本上可以被看做一个风险问题。它们到最后都面临着在多种（不确定）因素的情况下选择并确定哪个选项最合适的问题，这些因素包括组织政策、数据收集、成本等。因此决策工具和解决这些问题的自动化设备是必需的。
- **数字对象的显著特性**：这是了解一个对象并长期地保持其可用性与真实性所必需。因此不但要研究怎样捕获这些特性，还要研究怎样保存这些特性和它们之间的关系。此外，显著的特性令保存活动变得可测量了。
- **交互性**：已经有大量的格式和格式类型存在了，每天都还有新的出现，而所有现存的解决方案都只关注这其中的一部分。有一些资料库能处理任何类型的数字编码数据，但很多组织和机构往往仍必须使用某些方法来解决格式问题。因此，数字资源长期保存需要在不同的服务提供商间建立交互性和信任。
- **自动化**：长期保存数字数据的过程包括几个步骤，这些步骤往往是手动执行的，其结果通常是集成的并作为后续步骤输入。因此流程的自动化是必不可少的。
- **语境**：即使数字对象往往只记载描述一个问题的单一方面的信息，它们出现时的上下文和语言环境也对长期保存十分重要。了解语境——如组织的政策、与其他对象间的关系等——对及时理解这个对象至关重要。
- **存储**：尽管保存数字数据的方法有很多，尽管在优化数据大小方面已经做出了很大的努力，存储容量的不足仍是个大问题。然而，这一领域对网格等类似的技术研究起到了重要作用。
- **实验**：在每一个学科中，实地试验都是了解用户的反应和对数字知识库的需求的唯一方法，因此测试平台和实验的设计都发挥了巨大的作用。

3.2PARSE.Insight 的路线图

该路线图为科学数据的长期保存基础设施所需的各组件和方面提供了一个概览。

- **金融基础设施**：数字资源长期保存中缺少金融方面的概念和元素，数字保存的商业模式因其长期性而与其他情况不同。一个稳定、强大、可升级的基础设施——如储存设施—

—需要建立这些商业模式和服务与组件的资助计划。

- **虚拟化的政策、资源和流程**：虚拟化是从底层实现隔离服务的一种常用技术，它提出了一系列的要求：服务间的解译标准；抽象服务（如存储）；存储资源的备份；资源、数据和用户的逻辑命名空间。
- **可共享的表示信息知识**：一个增强了的表示知识管理能为验证表示信息提供一个半自动化的方法或充足的信息，可共享的表示信息是这一服务的基础。
- **可共享的硬件、软件知识**：其目的是建立一系列能便于交换硬件、软件和技术过时退化信息的服务。
- **数字对象的真实性**：数字对象的真实性需要通用的形式、标注和政策来证明。不久之后，一个用户将可以通过这些来判断数字对象的真实性程度。
- **数字版权**：要有在不断变化和发展的环境中正确处理数字版权问题的能力。几个现存的法律制度不断地被改变，数字版权管理为长期保存计划和保存行动提出了很多课题。
- **知识库的认证**：用适当的工具和最佳的实践方案来创造一个数字资源长期保存系统的审核与认证程序。

3.3Dagstuhl 研讨会：数字保存的自动化

2010 年 7 月，该研讨会举办于 Schloss Dagstuhl 的莱布尼茨信息中心（Leibniz Center for Informatics），目的是查明数字资源长期保存领域新出现的问题并规划未来的研究和发展。下文中将对该研讨会上提出的研究论题进行一个总结，共分为七个主题。

- **预保存系统**：其基本思路不是去建立一个信息存储系统，而是将所有数字资源长期保存的方法整合进现有的系统中并将其转化为预保存系统，最终形成一个预防机制。
- **除了元数据以外**：工具和模型框架是必不可少的，它们能够捕获数字对象的预期和实际运用信息，能自动建立数字对象和他们所处流程的文档。
- **存储技术和协议**：这类主题不仅总结了数字对象的物理存储，实际上还包括了很多创新的想法和研究的难点，有从智能忽略和可证删除中自我修正并复制的编码还有 DNA 数据存储等新领域。
- **管理政策和法规**：这类主题侧重于政策的具体定义和各机构的数字保存政策、法规、指导方针之间的区别。进一步研究的难点是探索决策中的政策界限和权益人间的协议，还有不断变化的政策管理和不同数字对象的版本联合。
- **道德、隐私、安全与信任**：道德和隐私主题是与安全和信任紧密相连的，它们的相互作用带来了一些有趣的问题。
- **数字保存的评价和基准测试**：基准测试的缺失从根本上阻碍了现今的特性和质量保障技术的发展和完善。为了支持一些新方法的客观比较并量化现有技术的进步，带注释的基准测试数据是必不可少的。
- **应用范围**：以电脑游戏为例来研究这些应用，如果能有一个长期保存电脑游戏的方法的

话，其他数字应用也就很可能得到保存了。然而，这项任务涉及到硬件、软件、版权和法律等多个问题。

3.4 Wiki 平台上的 DP 研究

继 Dagstuhl 研讨会之后，从数字资源长期保存（DP）角度概括计算机科学研究领域难点促进了一个 wiki 平台的建立，该平台就这些新兴的难点开展了涉及全球读者的广泛讨论。这一首创计划是一个实际由欧盟资助但在形式上独立的项目。详见 http://sokrates.ifs.tuwien.ac.at/wiki/index.php.

3.4.1 基于基础设施的研究重点

近期业内人士分析表明，数字数据的生产速度首次超过了全世界存储容量的增长速率。这一事实与长期访问这些数据的需求相结合，将使数字信息生命周期产生两个需求：

- **存储优先化的自动与半自动技术**

这将成为决定那些数据应该被保存、哪些应该丢弃的必备技术，且由于数据庞大的数量，自动决策技术也是必需的。对于那些应该保存的数据，人们应该判断它们是需要中期保存还是重要到需要放在长期保存档案库中。其他临时数据（如 IP 数据包、视频监控摄像镜头等）也需要不同的保存策略。

- **价格低廉、按需存储和处理能力**

对于非临时性数据，可能需要强大的处理能力来对其进行归档（例如要求媒体常态化时就需要把文件格式转化为易于存储的 PDA/A 格式）或者数据管理（支持高效检索的索引、元数据提取、和语义浓缩）。这些需求都需要以新的方式应用已建立的网格技术（提供分布式存储和计算的基础）以及经济的云计算方法（提供价格低廉、按需访问的虚拟计算资源）。由存储机构和商业权益人提供的长期保存技术基础既昂贵又缺乏效率，相信数字资源长期保存会发展成为由专家为这些机构提供的一项外部服务或基础设施。数字保存仍然是图书馆员和档案工作者的专业领域，但他们不需要是大型数据中心的管理人员和技术专家。设想一系列由商业组织提供的全球存档服务（如亚马逊 EC3/S3 服务）和一系列政府组织提供的全球存档服务（如为保存公共产品和文化遗产提供基础设施的计算机中心），这些服务将包括采集和访问，当然也要具备避开昂贵的数据传输，直接在存储系统执行保存操作（特征提取、索引、按需转化格式等等）的能力。这种方法也需要运用标准的 API 进行效用计算、虚拟化和资源配置，如正在开发的云社区和网格社区。

虽然在过去欧洲的研究曾假设比特流的保存问题已经得到解决，但事实并非如此。硬件系统、存储媒介和人工操作的不可避免的失败会永远影响字节的长期完整性。因此，全球保存服务必须基于有高复制性和冗余度的分布式系统，有巨大可扩展性的文件系统（如 ZFS）和嵌入式固定性测试还有点对点数据完整性检测都是必须要研究的。

分布式计算和存储网络间的交互性将是另一个难题，特别是这些网络中的节点中存在大量的任意时间内的遗留系统。最后，为了确保外部保存服务所需的信任水平，需要进行网络

安全领域的研究，有关知识产权、版权和隐私的政策法规也必须被考虑到。

3.4.2 基于内容的研究重点

过去的项目很大程度地把重点放在以文件为单位的非结构化内容上（包括很大一部分非结构化的科学数据）。今后的研究也应考虑结构化数据，尤其是数据库。其他欧洲行业需要重视的技术内容包括建筑和工程数据、药品数据和医疗记录。此外，软件尤其是开源软件档案的保存将成为一个重要的政策性问题。从技术上来说，这不一定被迫切需要着，但因为开源转换器在未来的存档资源检索中非常重要，所以这必须是一个优先考虑的主题。

除此之外，在访问存储对象、转换器和应用程序中新型的数字对象将越来越重要，如机器制造的虚拟图像。所以必须研究这些虚拟图像的开放标准化和为虚拟图像创造独立硬件的方法。

4 总结

该报告对欧盟在 ICT 计划下资助的项目中的数字资源长期保存研究活动做了一个概述。共分析了 18 个已完成或正在进行的项目，并明确叙述了其各自的主要目标。

编译自：

http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf

（李丹丹　吴振新校对）

【重要文献摘译】

《分布式数字资源长期保存指南》摘译之六

第六章 PLNs 的内容摄取、监管和恢复

Katherine Skinner，Matt Schultz 著 何莉娜 编译

概述

本章详细说明了 PLN 环境中内容是如何摄取、监管和恢复的。这些技术方案对处理和长期保存机构提交到 PLN 中的内容十分重要。它们是内容提供者和摄取并维护资源复本的其它保存网络成员的共同责任。

这些方案要么由PLN成员（如MetaArchive Cooperative）管理，要么由LOCKSS成员（ADPNet）管理。不同的PLN体制中，用于完成这些任务的工具也是不同的。本章主要介绍了MetaArchive Cooperative在其自管理的网络环境中所使用的程序和工具。下面所介绍的所有工具都是可免费获取的开源组件，它们都是基于LOCKSS基础设施的标准配置，或者是与LOCKSS系统协同工作的模块。本章总结了MetaArchive Cooperative的大量工作和基于五年中运营PLN的经验所提出的许多建议。

内容摄取

本部分详细说明了将内容摄入到 PLN 之前所需准备工作的最后阶段，即适当的插件开发和测试、试验缓存或网络的使用、利用概观工具建立资源的题名数据库词条。之后，本部分又描述了指定用于摄取内容的缓存、提示那些缓存内容已准备好待摄取以及在本地缓存摄取内容的过程。

插件开发和测试

必须精确定义插件的参数，确保其可以指导 LOCKSS 后台程序摄取内容提供者打算保存的特定和所有内容。如果插件有错误，那么后台程序将不能够爬取内容，或者更糟糕，即爬取和摄入错误的内容。这样的情况很不理想，尤其是自从 LOCKSS 不再提供删除网络中内容的核心方法之后。因此，测试插件是内容提供者摄取工作流中的一个核心程序。

最初，MetaArchive Cooperative 鼓励参与机构的开发人员利用其自己的工具和程序开发和测试他们的插件。MetaArchive Cooperative 维护大量的插件仓储库，允许内容提供者自己管理测试流程。这种方法产生了不同的结果：有些内容提供者为网络生产了固定的插件，有些生产了错误的插件。由于近两年来 Cooperative 致力于标准化其运营操作，其决定通过提供插件模板、开发文档、利用分散开发的核心插件仓储库、测试和发布程序来尽可能地将编

写和测试插件的过程标准化。这个趋势简化了内容提供者和网络层面的插件创建。内容提供者依然负责插件的编写和测试，但是目前主要由其中的一个中央基础设施来做这份工作。

基于此经验，Cooperative 建议其他独立的 PLNs 执行一个插件仓储库，将其以某种特定版本形式置于可获取的中心位置以便于开发和测试过程。同时，Cooperative 也建议网络创建标准的测试程序，尤其是在下述的测试网络中。

测试网络最佳实践

插件只有经过内容提供者的测试并且被认为产出达到了预期的爬取结果之后，才可以投入网络中使用。测试网络可以在本地或中央实施。在本地环境下，每个内容提供者使用一个题名数据库的临时实体，在 LOCKSS 后台程序上运行一个手工命令以测试插件。网络可以选择将测试网络作为供所有成员使用的核心资源。之后，内容提供者可以在其测试缓存上通过测试网络的缓存管理器或 LOCKSS 后台程序的用户界面来验证插件的精确性。

变更资源的再测试

当内容提供者对摄取后的内容进行结构上的变更（包括内容存储结构的改变）时，必须确保新的或变更后的内容可以适当地摄入到网络中。修改后的内容必须适合之前定义的资源参数，或者重新定义插件的资源参数以覆盖新的或修改后的材料。

如果资源的目录结构或位置改变了，内容提供者可能需要创建新的存档单元或修订资源的插件以提供适当的参数和存储单元，帮助后台程序继续在一定的间隔时间段再次摄取内容。

任一原因的插件变更都需要在测试缓存或测试网络上进行新一轮的测试。

定型题名数据库

一旦测试摄入得到验证，内容提供者必须告知其它 PLN 其内容已经准备好并且可以摄入到成果网络中了。独立的 PLNs 可能使用 MetaArchive Cooperative 开发的概观工具促进这一过程。

概观工具

概观工具是一个基于网络的数据管理工具，其同时维护关于提交摄取的每个数字资源的特定 LOCKSS 技术元数据和描述性的资源层面的元数据。内容提供者可以使用概观工具创建、更新、维护他们的资源描述。概观工具同样生成特定的 LOCKSS 配置数据，这些数据存储在题名数据库中，系统程序利用这些数据来配置网络。概观工具包含一个由 Cooperative 设计的元数据框架，该框架包括广泛使用的元数据标准（如 Dublin Core、METS 和 MODs）的一些元素。

内容提供者在概观工具中为每个资源准备一个资源描述说明，为资源提供一个名称和标题，并录入资源层面的元数据。内容提供者同样录入资源配置参数（关于存储单元和插件的信息）。内容提供者同样检验资源、元数据及其摄取信息，确保内容和其语境被适当地定义用于长期保存。

指定摄取缓存

一旦内容提供者指定了待摄取的资源，其就会将经过测试和验证的插件版本移动到插件仓储库的已发布程序中，以此发布资源的爬取程序。由于 PLNs 配置了一个分布式的基础设施，其中每个缓存都是独立自治的（如，即使是中央成员也无权直接访问每个缓存），没有直接的方法来初始化网络中缓存的内容爬取。因此，每个 PLN 必须建立一个主要机制，当新资源准备好待摄取时，提醒缓存进行摄取。如果网络太大，并不需要每个缓存都摄取内容，那么 PLN 同样需要决策怎样只提醒那些被指定收割内容的缓存。

通讯最佳实践

项目邮件服务和电话会议可以为指定站点复制和通知摄取过程提供有效的警示策略。例如，在 MetaArchive Cooperative 环境下，缓存的任务和摄取信号是通过下面一系列的综合事件完成的：（1）内容提供者通知 PLN 中央系统管理员资源已准备好待摄取；（2）中央系统管理员核实内容提供者已经成功完成了的摄取测试和题名数据库概观工具条目，要求 PLN 项目管理人员指定 7 个站点摄取内容。站点的选择主要基于缓存的容量、机构贡献的公正量与站点本身拥有的公正量比较以及地理上的分布；（3）项目管理者和中央系统管理员核实这些存储站点；（4）项目管理者利用概观数据库正式指定这七个站点。项目管理者同样通过项目邮件服务联系所有指定的缓存，通知其有新的资源需要爬取；（5）指定成员的缓存管理者将已鉴定的资源通过他们的 LOCKSS 后台用户界面加入到其缓存配置中，以启动爬取进程，同时监管爬取过程保证爬取的顺利完成；（6）PLN 的中央系统管理者、内容提供者和指定的缓存站点联合负责，以保证资源得以适当地爬取和成功地摄入，所有各方需要在每周的电话会议上做报告，核实其成功地摄取了资源，或讨论缓存中可能出现的问题。

内容一旦被摄入网络中，LOCKSS 软件就会积极开始评价和长期保存那些内容，下面将会详细阐述这一问题。

内容保存

本部分介绍了缓存处理内容完整性和正确性的过程、活跃内容和封闭内容的区别、为有效的网络保存配置插件和站点服务器的重要性。

轮询

当所有的指定缓存完成了摄取过程，这些缓存会定期参与轮询比较所摄取的文档，确保内容的复本相匹配。

在成功的轮询中，如果某一缓存中的内容和其它不匹配，LOCKSS 后台程序会识别不一致的缓存，启动对内容的重新爬取，从内容提供者的站点上爬取公认的权威复本。如果内容提供者站点不再可访问，不一致的缓存内容由其它保存缓存修复。

活跃资源和封闭资源

内容提供者对他们所提交的每项内容的长期保存活动都有两种选择。他们可以要求 LOCKSS 后台程序积极回访摄取站点的复本，或者从分段服务器上一次性摄取内容，并指定资源是封闭型的、不可再次摄取。

活跃模式

默认情况下，在网络缓存上运行的LOCKSS软件执行常规的、随机的爬取和轮询机制，在此过程中，其利用插件和题名数据库中记录的基本URLs和爬取资源的规则，通过HTTP或HTTPS反复摄取内容。通过这种活跃模式摄入到PLN中的内容，应该仍然可以通过web来访问。活跃模式允许保存网络进行文档更新，并且如果爬取参数合适，允许将新文档添加到内容数据库中。同时，该模式也允许内容已毁坏的缓存从内容提供者站点上的权威信息源处重新摄取内容。

封闭模式

某些情况下，PLNs可能选择摄取那些通过网络不再可获取的资源。内容提供者可能会避免存储受禁的内容、或者网络可访问服务器上受版权保护的内容。在这种情况下，内容提供者会将其内容临时迁移到一个分段服务器上以待摄取。一旦内容成功摄入，并且所有的指定缓存都拥有同一的内容复本，那么就可以从内容提供者网站上撤走这些内容，而保存的复本即成为权威的内容版本。如果资源中的某一文档损坏了，网络在其常规轮询中会发现这一问题。由于缓存可以比较复本，它们将会达成协议，判定一组缓存相互匹配，规定缓存持有的复本就是权威的版本。其中的一个缓存将会修复不一致的缓存上的损坏复本。

重新爬取间隔

对于活跃模式下摄取的内容，内容提供者必须在其资源插件中设定一个合适的重新爬取间隔。大约一个月爬取一次任一站点（尤其是大型资源）对网络来说开销太大，并会减慢轮询和其它长期保存进程。目前最佳的实践说明，静态站点最多一个月需要进行一次重新爬取，最少一年需要进行一次爬取。重新爬取间隔必须同样在封闭资源的插件中设定——除非不需要重新爬取。

另一个值得独立的 PLNs 执行的就是最后修订的网络服务器配置能托管内容提供者的资源。这种设置使得 LOCKSS 后台程序可以比较存储单元中的 http 标题信息以判断参与站点服务器上的文档是否比缓存磁盘中的文档旧。之后，LOCKSS 后台就可仅仅摄取那些文档。这减轻了缓存和整个网络的负担。

内容监管

内容监管的责任由内容提供者、每个保存站点、各种指定中央职员（LOCKSS 团队或 PLN 托管的中央职员）共同负责。本部分描述了中央职员和成员机构的责任，以及目前 PLNs 可获取的监管工具。

职员和成员责任

内容提供者站点的缓存管理者有责任保证所有指定保存站点已经成功摄入了其所有内容。这要求内容提供者不仅能检验复制，而且要同时拥有一个访问复制缓存站点内容的代理和内容审计手册。

保存站点的缓存管理者有责任摄取和监管缓存中的内容。这使得当新内容可获取时就会被摄取，可以监管摄取程序以确保其完整恰当，监管网络缓存间在轮询和修复过程中的通信。

中央职员承担管理整个网络和监管其长期保存活动的责任。管理包括监管整个网络行为、监视系统记录、保证轮询和修复程序运行正确、确保网络运行平稳。

监管工具

有三种辅助内容监管程序的工具：缓存管理器、LOCKSS 后台用户程序、cron 报告：

缓存管理器是开源的、基于网络的工具，由 LOCKSS 团队和 MetaArchive Cooperative 共同开发。其关注个人缓存上的 LOCKSS 后台程序，收集与 AUs 保存有关的状态信息、整个网络状态以及缓存存储能力。

LOCKSS 后台用户程序是 LOCKSS 软件装置提供给所有管理缓存的 PLN 成员的一个核心工具。该程序罗列了缓存可摄取的存储单元，传达关于存储性能的深层信息，作为验证个人存储单元完整性的主要工具，内容丢失时可以进行内容的恢复。

Cron 工作可以被配置用以提供与 LOCKSS 后台程序进行的轮询结果以及缓存上存储单元状态有关的调度报告。

内容提供者站点的缓存管理者可以利用缓存管理器观察哪些地理上分散的站点已经复制了其内容，以验证内容是否被成功摄取。管理缓存者以及任何中央 PLN 职员同样可以利用缓存管理器监管缓存是否适于摄取新内容。

通过 LOCKSS 后台用户程序，内容提供者站点的缓存管理者或任何中央 PLN 职员成员同样可以通过代理定期审计内容。这一特征帮助确保了插件中的收割参数仍然能够指导 LOCKSS 缓存正确地摄取和更新内容。缓存中的代理资源在标准网络浏览器中的 URL 中变得可见，并且可以手动查看以保证其完整性和正确性。

最后，中央 PLN 职员成员可以在某一时间启动 cron 工作自行运转，或更新、报告 LOCKSS 后台程序在缓存间进行的轮询结果。这记录了网络和缓存上内容健康状况的实时信息。可以对这些报告进行配置并通过邮件传送给必要的人员，如果有需要，同时传送给内容提供者站点的技术和管理人员。

恢复内容

当内容提供者站点发生灾难性损毁时，可以从另外的拥有内容复本的缓存将内容恢复。相似地，当保存站点发生灾难性损毁时，可以通过再次重新摄取内容提供者站点或网络中任何其它包含有相关资源的缓存中的内容而得到修复。

恢复内容提供者站点

恢复参与机构遭受重大灾难性损毁的内容需要在网络中的其它缓存（这些缓存专门用于复制内容）中检索其长期保存的复本。

需要访问保存在缓存中的内容的参与机构，利用网络浏览器请求一个或多个保存的 URLs，通过 LOCKSS 后台用户界面的审计代理功能分布获取一个代理 PAC（Proxy Auto-Configuration）文档。通过这一装置启动抽取，可以精确地检索到原始摄取或重新爬取的整个资源。具备这些修复的内容碎片，同时参考题名数据库的资源实体，成员应当有能力在一定的时间范围内在新的或修复的网络服务器上重建资源。

恢复缓存

缓存失效时，参与机构在网络上建立新的缓存。新缓存从指定摄取内容的站点上重新摄取资源。通过网络不可获取的资源（如上述的指定封闭资源）可能从其它拥有那些内容复本的缓存上修复到新缓存中。

附加说明

LOCKSS 软件保证与插件中收割参数相匹配的新增文档和修订文档，将被网络抓取为独立的个体文档版本。LOCKSS 无法自主判断内容提供者站点的变更是否经过授权，因此它会将其所收集的所有文档都作为权威的文档对待。LOCKSS 收集到之前已经摄取了的文档的话，它不会重写现存的文档，而是建立一个新版本——因为内容提供者可能最终需要恢复最初的版本，该版本更能反映权威的内容。

这些版本在成员的 LOCKSS 后台用户程序中是可见的。如果内容提供者遭受灾难性事件而需要利用保存的复本修复时，所有存储的版本都可通过该程序轻松获取。

结论

一旦内容成功摄入 PLN 中，所有成员必须完成其各自的以及共同的责任，保证内容得到了妥善保存。分布式长期保存的目标就是在发生损毁事件时提供修复的权威版本。在 PLN 环境下达到这样的目标，需要内容提供者、缓存管理人员以及中央职员在保存过程的每个阶段都积极参与。

编译自：http://www.metaarchive.org/GDDP

（李丹丹 吴振新校对）

【动态追踪】

JISC 学术研究完整性会议

随着高校和科研人员保存研究数据的需求日益迫切，JISC 学术研究完整性会议仔细考虑了大学在保护学术研究完整性中所起到的作用，并从战略和技术的角度调查了它们所面临的实际问题。

研究人员及其所属机构越来越多地被要求长期保存他们创建的研究数据以便重复利用，否则将会破坏公众对科学的信心，近期的一些引人注意的案例正说明了这一点。然而，良好的科研数据管理并不容易实现，一方面它对科研人员的传统工作方式有一定的影响，另一方面，机构相应的科研支撑基础设施需要落实到位。

会议包括以下主题：

（一） 机构对于良好数据管理的思考

（二） 支持优秀调研

（三） 挑战与应对

A. 共享研究数据的法律义务与政策责任

B. 机构责任

C. 公开发行纪录的完整性

（四） 从国家战略高度对良好数据管理的思考

会议相关视频资料及其他资源网址：

http://www.jisc.ac.uk/events/2011/09/researchintegrity/resourceshub.aspx

编译自：http://www.jisc.ac.uk/events/2011/09/researchintegrity.aspx

（么媛媛编译，齐燕 吴振新校对）

翻译的数字化实践：OAAP 项目的启示

OAAP（Our Americas Archive Partnership）项目包含着学者、图书馆员和信息科学家们的共同努力，它为发现、获取、使用存在于多个数字仓储中的学术著作提供了一个集成的方法。该档案库由最初著于1492年到1920年间或关于此期间的美洲的电子文本和图像构成。该项目的目标是全方位地呈现一个复杂的、多语种的美洲，并从半球的角度促进关于美洲文化和历史的新研究。这篇文章讨论了将多种语言的历史文件数字化过程中的复杂问题，包括创造一种原生数字化（born-digital）翻译方法、生成能够最确切地描述这些稀有且原始的文件

的特定的元数据。

OAAP的译文为更好地访问非英语语言的重要一手资源的内容提供了便利。它们为针对美洲各国和半球文化的新研究和新知识的出现提供了有力支持。这些资源可以看做是探索原生数字化翻译这种数字处理方法的一个语料库，这种翻译方法包括进行基于史实的注释、对文本进行语义编码和以有意义的形式来描述这些资源。

编译自：http://www.dlib.org/dlib/september11/rivero/09rivero.html

（么媛媛编译，齐燕 吴振新校对）

SDSC 推出可升级、高性能的云数据存储服务

2011 年 9 月 22 日，加利福尼亚大学圣地亚哥超级计算机中心（SDSC）宣布美国最大的学术云存储系统上线。该系统被认为是专为研究人员、学生、学者以及业界用户量身定做的，可为用户提供稳定、安全、经济的数据（尤其是大数据集）存储和共享服务。

SDSC 主任 Michael Norman 表示：“我们相信对于广大的研究人员，SDSC 的云服务会很好地变革数据存储和共享的方式，尤其是对于越来越流行的大数据。”“为了能够达到联邦数据分享的要求，SDSC 云存储做出了很大的努力。我们要让每一个数据对象都有一个属于自己的 URL，并且可以通过互联网访问。”

SDSC 基于网络的新系统是 100%基于磁盘的，并且通过最高速度 10GB 的交换机连接，具备高速的读写性能。目前 SDSC 云存储系统的初始容量为 5.5PG（相当于 2 亿 5 千万页的文本），而其每秒的数据读取速度可达到 8-10GB，随着更多的节点的加入以及容量的扩充，这一速度将会不断得到提高。此外，通过聚合性能和容量的线性增长，SDSC 云存储系统的数量级可以从几百字节扩展到千万亿字节。

SDSC 的云存储系统的上线运行将惠及的用户和研究伙伴包括：加州大学图书馆、医学院、Rady 管理学院、Jacobs 工程学院、SDSC 研究人员，同时还包括由美国国家科学基金会、国家卫生研究所和医疗保险服务中心共同资助的研究项目。

正如 SDSC 副主任 Richard Moore 所言：“SDSC 云服务标志着我们对待长期保存的思考模式正发生转移。我们正在从档案数据‘只重保存，不重获取’的模式转化为当下的模式——正如大家所说的‘如果你认为你的数据很重要，那么它们就应该在更广的范围内被很容易地访问和共享’”

编译自：http://www.sdsc.edu/News%20Items/PR092211_sdsccloud.html

（刘晓敏编译，么媛媛 吴振新校对）

FDA 存储库数据量超过 100TB

佛罗里达数字档案中心（FDA）创办于 2005 年 9 月，由美国图书馆服务协会拨款，由佛罗里达自动化图书馆和数字化档案馆中心（FCLA）负责开发。FCLA 工作人员为 FDA 专门设计开发了 DAITSS(Dark Archive in the Sunshine State) 系统，在 2005 年 9 月发布了 DAITSS 早期版本，又于 2006 年发布了最终的完整版。通过提供一个高效率的长期保存贮存库，FDA 可以为佛罗里达州的教学、研究活动提供支持。

截止至 2011 年 9 月 21 日，佛罗里达数字档案中心（FDA）存储库数据量已经超过 100TB（一万亿字节），其中包括 299,415 个 AIP 和超过 4 千 3 百万份的文件。每个 AIP 都有单一的副本，所以 AIP 的总拥有数据量超过 200TB。

编译自：**http://fclaweb.fcla.edu/content/fda-repository-now-holds-over-100tb-data**

（刘晓敏编译，么媛媛 吴振新校对）

【信息扫描】

POCOS 格拉斯哥座谈会－－软件艺术

软件艺术是一种正在迅速成长的艺术发展类型，它已经引起了来自于艺术界和文化事业单位的浓厚兴趣。目前有很多软件艺术作品已经在全球重要的博物馆中展出，因此对它们进行获取管理、长期保存迫在眉睫。对这种基于软件的艺术作品的长期保存面临诸多挑战，包括：对象间拥有着复杂的相互依存关系；需要有时间、交互的考虑；开发技术和方案的多样性。

2011 年 10 月 11 日-12 日在格拉斯哥召开的为期两天的软件艺术研讨会将提供一个平台供与会者讨论这些挑战， 审查和讨论这个领域的最新发展，分析现实生活中的相关案例，并从事各种交流活动。研讨会将重点讨论以下关键问题：

- 软件艺术长期保存的涵义和进展
- 易逝性的问题
- 软件艺术作品的重要特性
- 与一般软件的长期保存的关系
- 软件艺术长期保存的实践

相关链接：http://www.pocos.org/index.php/pocos-symposia/software-art

编译自：

http://www.openplanetsfoundation.org/events/2011-10-11-pocos-glasgow-symposium-software-art

http://www.dcc.ac.uk/events/

（齐燕编译，刘晓敏 吴振新校对）

美国国会图书馆开展数字保存拓广与教育计划

数字保存拓广与教育计划基层研讨会 9 月 20 日至 23 日在华盛顿特区图书馆举行，有来自全国范围的 24 位在职的专业人士参会。这次研讨会以一种研讨会模式试运行，目的是为了培养国家的教师团体，来教授人们有关数字资源长期保存的基本原则和实践。他们都至少有一些社区培训和数字保存两方面的经验。

该研讨会受到美国国会图书馆开展数字保存拓广与教育计划的支持，任务包括推进美国国内数字保存的拓广和教育，鼓励更多的个人和组织积极地保存他们各自的数字内容。它致

力于在美国全境内分享优质的数字保存培训经验，最终使任何对此有兴趣的组织或个体都可以获取并支付的起相关费用。

该研讨会的独特性在于，它是专门为将要从事数字保存的实际从业人员设计的。而且，这个研讨会不是针对管理者，而是面向新手从业者。它也试图努力做到开放和低成本。并希望通过这一事件能够促进全国的开放的易用的数字保存培训。

编译自：

http://blogs.loc.gov/digitalpreservation/2011/10/graduates-to-sow-seeds-of-new-training-program-across-u-s/

（齐燕编译，刘晓敏 吴振新校对）

JISC 服务教学、科研的云计算——已发布的项目和合作伙伴

自从二月份宣布成立一个 £1250 万的基金来帮助大学和学院通过更有效率的协同合作来创造更大的经济价值以来，HEFCE 和 JISC 现在能够确定所涉及的项目和合作伙伴来完成以下两个部分的工作：国家级的云计算基础设施以及相应的配套服务。

英国联合科研网将为高等教育机构和商业供应商间的云服务的采购事宜提供国家级的经纪业务；Eduserv 为高等教育机构提供一个云计算基础设施试点。

其他合作伙伴包括德蒙特福德大学、埃克塞特大学、爱丁堡大学、肯特大学、利物浦约翰摩尔大学、牛津大学、莱斯特大学、南安普顿学院和桑德兰大学。数字管理中心(DCC)将开发数据管理工具和培训项目。这将会支持各种数据管理计划的制定和实施，来帮助高校及其研究人员进行数据的长期保存以备共享、重用和引用。

编译自：

http://www.dpconline.org/newsroom/whats-new/752-whats-new-issue-37-september-2011

（齐燕编译，刘晓敏 吴振新校对）

VIDaaS 项目启动

基于牛津大学的计算服务，VIDaaS(**Virtual Infrastructure with Database as a Service**))项目于 2011 年 5 月正式启动。它有两个重要的、互相联系的目标。首先，它将为学术研究者开发一种网络数据库服务(DaaS)，使得用户能够在线地创建、使用和共享各种数据库。这种方式有以下几个优势：自动安全备份、能够随时随地存取数据、允许多个合作者同时使用同一个数据库。对于有需要的用户，它还将提供在网上发布数据集的一种简单方法。其次，该项目将搭建一个虚拟基础设施，使 DaaS 功能在云计算环境中运行。云技术——即使用一

个由各种计算资源组成的网络来存储和处理数据——通过规模经济和灵活性的提高，很有潜力为组织节省数目可观的成本。

编译自：

http://www.dpconline.org/newsroom/whats-new/752-whats-new-issue-37-september-2011

（齐燕编译，刘晓敏 吴振新校对）

21 世纪长期保存路线图：迈向数字未来

2011 年 10 月 20 日，美国国会图书馆举办了一场题为“二十一世纪的长期保存路线图：迈向数字未来”的研讨会，旨在解决文化遗产机构在平衡遗产收藏保护需要与日益复杂的转换和原生数字遗产的收集要求时所面临的各种挑战。

研讨会发言人包括来自美国国家档案与文件管理局、史密森学会、美国国家公园管理局、美国国会图书馆、图书馆和信息资源委员会和一些的基金会的高级管理人员。

藏品和项目的管理人员、开发人员、负责任和赞助者受邀参加此次会议。会议日程和登记信息详见：:www.loc.gov/preservation/outreach/symposia/transition/.

编译自：http://www.loc.gov/today/pr/2011/11-183.html

（齐燕编译，刘晓敏 吴振新校对）

第四届 eSciDoc Days 会议召开

由马克思.普朗克数字图书馆（Max Planck Digital Library）和卡尔斯鲁厄专业情报中心（FIZ Karlsruhe）共同主办的第四届 eSciDoc Days 会议于 2011 年 10 月 26 至 28 日在柏林举行。eSciDoc 是由马克思.普朗克数字图书馆和卡尔斯鲁厄专业情报中心共同开发的一个开源 eResearch 环境，通过相关元数据、联系以及访问权限，为研究出版物数据管理提供应用和服务。

本次会议的主要内容将集中在协作化 eResearch 环境及其面临的相关挑战，如：如何对在研究过程中持续增长的数据进行可持续的管理并提供研究结果以及研究数据的出版环境。有不同专业学术背景的专家也将就数字信息基础架构的发展近况进行讨论。

本次会议的主讲嘉宾是来自英国联合信息系统委员会的项目经理 Simon Hodson。会上，Simon Hodson 将与大家共同分享他在推动和支持高校教育与科研的数据管理和数据共享方面的技术和知识。

会议日程安排：

第一天（10 月 26 日）

对 eSciDoc、eSciDoc 社区及其基础服务和定制化应用作一个总体回顾；

第二天（10 月 27 日）

主要围绕三个焦点主题：“基础设施项目（Infrastructure Projects）”、“研究型数据管理（Research Data Management）”和 “eSciDoc 入门指南（eSciDoc - Getting started）”。

每一个议题都将提供最新的进展报告，并通过相关的实践综合介绍 eSciDoc 及其应用领域。此外，还将为具有技术背景的与会者提供两个额外的 eSciDoc 教程，方便他们熟悉 eSciDoc 基础架构并开发出属于他们自己的 eSciDoc 应用。

ESciDocDays2011 网址：

https://www.escidoc.org/JSPWiki/en/ESciDocDays2011Program

编译自：https://www.escidoc.org/JSPWiki/en/Startpage_blogentry_040811_1

（刘晓敏编译，么媛媛 吴振新校对）