2012 年第 4 期（总第 14 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

2012 年 4 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

【信息扫描】

一战期间诗歌的数字存档

第一次世界大战诗歌数字存档是一个拥有7,000多个供教学、学习、研究使用的文本、图片、音频、视频文件的在线知识库。该多媒体数据库由初级的原始材料（如诗歌手稿、书信、日记等）加上背景信息（帝国军事博物馆的图片、音频和影像资料）组成，可浏览、检索及免费获取。

该存档库的核心内容包括该时期主流诗人的重要资料，如Wilfred Owen，Isaac Rosenberg，Siegfried Sassoon，Robert Graves，Vera Brittain，Edward Thomas，Roland Leighton，Ivor Gurney，Edmund Blunden和David Jones等。包括战争中和战后回忆时他们写的诗歌的各种形式的手稿、日记、信件和服务记录。通过利用来自帝国军事博物馆数字化的同一时期的艺术品，该档案馆扩大了其职权范围，包括关注战争期间和帝国军队中妇女发挥作用的相关资源和在大后方创作的作品。

该档案库还包括从帝国军事博物馆补充的大量多媒体艺术品，大众贡献的超过6,500条的独立档案，和一套专门开发的包括在线教程、教学包的教育资源，还有一个在三维虚拟世界Second Life（第二人生）中的展览。

编译自：http://www.dlib.org/dlib/january12/01contents.html

（么媛媛编译，李红培 吴振新校对）

Alfred P. Sloan 基金会资助 NISO 和 OAI 进行资源同步标准的设计

国家信息标准组织（National Information Standards Organization，NISO）和开放档案协议（Open Archives Initiative，OAI）的一个联合项目得到了二十二万两千美元的拨款，该项目旨在为网络资源的实时同步开发一个新的开放标准。越来越多的大型数字馆藏可以从多个主机位置获取，在多个服务器上进行缓存，并利用多个服务。这些作品或数据的副本在网络上的扩散使得这些知识库保存其藏品以及提供及时准确的服务越来越困难。当从一个文件网络转移到一个数据网络时，同步就变得更加重要：基于不同步或不连贯科学或经济数据而做出的决定可能会有很大的负面影响。

据 Memento 项目的主要研究人员、美国欧道明大学(Old Dominion University)副教授

Michael L. Nelson 介绍：这项建议针对 Memento 项目环境中暴露出的问题开发了一个协议，在网络上使用有统一的访问时间戳的资源版本。利用 Sloan 的资助，他们组建了一个一流的核心团队来制定标准。团队目前已经做了一系列信息互操作方面的工作，如 Memento；OAI 对象重用和交流（Object Reuse and Exchange），用于描述网络资源聚合的协议；开放注释（Open Annotation），以资源为中心的注释框架； DSNotify 为关联数据的变化检测框架。

康奈尔大学信息科学副教授、OAI 主管 Carl Lagoze 认为：元数据获取协议 OAI（PMH）2.0 可以用来有效地同步资源的元数据，但同步的资源本身从来都是不确定的。虽然有一些资源同步方法，但它们一般都是专门的、由参与其中的个人来安排的，无法广泛应用。

洛斯阿拉莫斯国家实验室科学家、OAI 主管、Memento 项目的主要研究者 Herbert Van de Sompel 认为：网络基础设施缺少一个互操作的、高效的、轻便的机制来支持大规模同步资源，当开始研究这一问题后，这种情况变得越发明显，有一些能帮助解决现有问题的候选技术，但需要将它们放在一起并以适当的方式进行评议。

NISO 总经理 Todd Carpenter 希望这个新标准可以帮助知识库管理者通过复制和更新过程的自动化来节省大量的时间、精力和资源。最终的结果将是增加网络知识库内容的总体可用性，减轻当今互联网上的过时、不准确、废弃内容带来的一系列问题。

编译自：

http://www.niso.org/news/pr/view?item_key=238435856f6666cad3d0bf3a3eb70caa310e1e3c

（么媛媛编译，李红培 吴振新校对）

DAITSS 保存仓储软件发布开源版本

由弗罗里达图书馆自动化中心（Florida Center for Library Automation，FCLA）开发的 DAITSS（Dark Archive in the Sunshine State）软件根据 GPL v3 许可可以免费获取。

DAITSS 是由弗罗里达图书馆自动化中心在 IMLS 支持下开发的一套用于数字资源长期保存的仓储软件。目前弗罗里达数字存档项目使用了 DAITSS，该存档是 FCLA 提供的一项长期保存知识库服务，供弗罗里达的十一所公立大学图书馆使用。DAITSS 是在 2005 年开始开发的，最近重新改写了系统架构，更易于实施、维护，具有更强的可测量性和可扩展性。

DAITSS 为提交、摄取、档案存储、访问、提取和知识库管理功能提供自动化支持。它提供了一套 RESTful 网络服务和微服务，执行了严格的管制以确保存档内容的完整性和真实性。DAITSS 实施了基于特定格式处理的主动保存策略，包括必要时进行规范化和前向迁移。它特别适用于文本、文档、图像、音频和视频格式的材料。

DAITSS 适用于多用户环境，支持图书馆联盟和机构保存仓储。

编译自：http://daitss.fcla.edu/content/press-release-daitss-open-source-site

（么媛媛编译，李红培 吴振新校对）

第二届数字保存 LIBER 国际研讨会

欧洲数字资源长期保存中的合作伙伴关系

LIBER 文物收藏和保护督导委员会与佛罗伦萨数字文艺复兴基金会、荷兰国家图书馆在意大利佛罗伦萨合作举办了第二届数字保存 LIBER 国际研讨会。

该研讨会对一些最知名的合作行动进行了概述：利益相关者的参与、基本的法律基础、商业模式——并帮助与会者分析适应其组织、资源类型和国家文化背景的最适合的选择方案。会上讨论了影响与会者选择决策的组织问题、法律问题、财务问题和技术问题。同时，研讨会上还展示了一些最佳实践。

编译自：

https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1201&L=digital-preservation&F=&S=&P=19241

（么媛媛编译，李红培 吴振新校对）

【动态追踪】

DuraCloud Wiki 网站开展新的研究项目

DfR（DuraCloud for Research）是一个基于云技术的开源数据存储和管理项目，旨在满足科学家、研究员和数据管理者的工作过程中的特定需要。该项目已在DuraSpace wiki建立了一个网站，网址为 ：

https://wiki.duraspace.org/display/DURACLOUDDTR/DuraCloud+for+Research

该网站包括几个栏目：

- 项目介绍——提供了整个DfR项目的介绍
- 项目状况——包括项目开发进度的概述，提供对相关DfR研讨会报告和架构的链接
- 加入其中——列出了访问者可以参与回答的与DfR有关的问题
- 保持联系——加入DfR的邮件列表来获取其新闻和信息

- 资源——相关的文件和信息

DfR致力于为那些没有IT人员提供帮助的科学家创造一个安全、可靠、灵活且易于使用的管理研究数据的方法。有了DfR，从源头“抓取”研究结果意味着科学家和研究人员有能力接受授权机构对其保存数据和创建对数据的访问这两方面的问责。

DfR是DuraSpace的DuraCloud数据管理和存档服务的一个扩展，并且得到了Alfred P. Sloan基金会的赞助。目前项目正在开发中，新服务有望在2013年面世。

编译自：http://duraspace.org/new-duracloud-research-wiki-web-site

（李红培编译，么媛媛 吴振新校对）

KEEP 项目发布仿真框架 2.0 最终版

欧洲项目KEEP发布了其最终版本的仿真框架（Emulation Framework，EF）。该软件允许用户利用仿真来访问旧的计算机文件和程序，并且不需要复杂的安装和配置过程。该软件是开源的，任何组织和个人都可以免费使用，下载地址是：http://emuframework.sf.net。

然而，问题是为什么要使用仿真技术呢？假如一个组织或个人正需要了解几年前的数字文件或应用程序，会发现在现在的计算机系统上，它们可能连运行和打开都很困难。随着时间的推移，旧的计算机被新的替代，出现了新的计算机体系结构，新的软件有了更新或升级后的版本，这些都给读取几年前的原始文件或应用增加了额外的复杂性。目前的计算机无法执行旧的程序，旧的文件格式很可能需要转化后才能使用。有了仿真技术，就可以重新创造出旧的计算机平台来运行程序，给用户的外观感觉和以前几乎一模一样。

然而，配置这样的平台是很困难的。仿真框架通过对配置过程的自动化处理，架起了技术和可用性之间的桥梁。EF2.0配备了一套基本的7个开源模拟器，这些模拟器能在x86、Commodore64、Amiga、BBC Micro、Thomson和Amstrad等模拟平台上运行。经测试，该程序至少可以支持30种不同文件格式的访问。这套基本的支持模拟器和软件可以根据用户需求进行扩展。

EF框架是KEEP刚刚发布的软件。它包含仿真的一切功能而且无需进行复杂的安装和配置。因为它将这些原本需要手动进行的步骤自动化了：

（1）确定需要读取的数字文件类型；

（2）寻找所需的特定软件和计算机平台；

（3）将需求与现有的软件和模拟器进行匹配；

（4）安装模拟器；

（5）配置模拟器、准备软件环境；

（6）将所选择的数字文件插入模拟环境中；

（7）控制模拟环境；

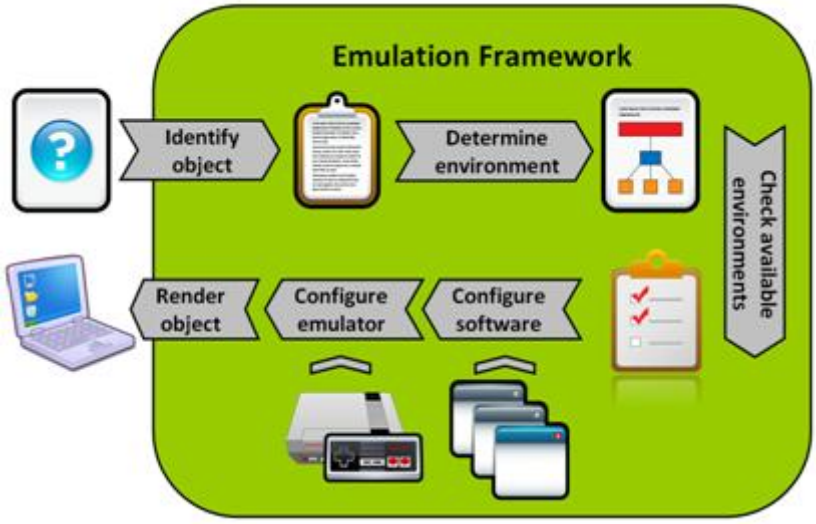

图1 仿真框架流程图

现在所有这些功能都被打包在一个易于安装的免费软件包中，可以在当前所有主要平台（Windows、Mac、Linux）上运行。

KEEP（Keeping Emulation Environments Portable）项目由2012年2月29日结束的欧盟第七框架计划资助。该项目开发一些服务来使静态和动态的数字对象得到准确的呈现，如文本、声音和图像文件、多媒体文件、网站、数据库、电子游戏等。它的总体目标是通过为广泛的数字对象的访问和存储开发灵活的工具，来为人类文化遗产的普遍访问提供便利。KEEP还解决了围绕仿真系统的媒介和专利/许可的版权保护的相关法律问题。更多的项目信息和成果参见：http://www.keep-project.eu。

EF软件开发团队由荷兰国家图书馆领导，主要开发由Tessella完成，包括设计其核心功能、软件存档和模拟器存档。初级用户界面由英国朴茨茅斯大学开发，高级用户界面由荷兰国家图书馆开发。

编译自：

http://www.openplanetsfoundation.org/blogs/2012-03-29-dream-perpetual-access-comes-true

https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1203&L=digital-preservation&F=&S=&P=7769

（么媛媛编译，李红培 吴振新校对）

迈向数字资源长期保存国际合作：iPRES2011 研讨会报告

“迈向数字资源长期保存国际合作”是2011年11月4日在新加坡举行的一个为期半天的互动研讨会，与IPRES2011相结合。该研讨会旨在推动跨国数字资源长期保存的集体行动。早先的ANADP（Aligning National Approaches to Digital Preservation）研讨会于2011年5月23-25日在爱沙尼亚首都塔林举行。该研讨会允许来自各种不同国家背景的个人共享关于建立支持数字资源长期保存的国际合作战略的信息和观点。

IPRES2011研讨会旨在为不在ANADP会议中的个人和机构开辟讨论空间，并对怎样制定具体步骤来推动国家间的联合活动进行重点讨论。

该研讨会吸引了来自10个不同国家的14名发言人。出席会议的有38人，来自14个国家：奥地利、加拿大、丹麦、芬兰、德国、印度尼西亚、日本、荷兰、新西兰、新加坡、南非、英国、美国。

会议涉及了一系列有关各种“联盟机会”的会谈，包括：

- 研究和教育——Özgür Külcü（哈杰特泰佩大学），Laura Molloy（格拉斯哥大学）和Andi Rauber（维也纳技术大学）
- 技术基础设施——Raju Buddharaju (新加坡国家图书馆管理局)，Steve Knight (新西兰国家图书馆)和Bram Van DerWerf (开放行星基金会)
- 提供专业指导——Masaki Shibata (日本国立国会图书馆) 和Daisy Selematsela (南非国家研究基金会)

分组讨论中涉及的方面包括：

- 访问联盟——由Jon Crabtree（奥德姆研究所）组织
- 经济联盟——由Neil Grindley（联合信息系统委员会）组织
- 政策和促进公共利益——由Seamus Ross（多伦多大学）组织

最后对可能即将会采取的步骤进行了讨论。

相关个人如有兴趣，可寻找有关研讨会的更多信息，或加入数字保存的联盟国家方法小组来讨论进一步的举措、思想和活动，请访问以下页面：

http://digitalcurationexchange.org/international-alignment

编译自：http://www.dlib.org/dlib/march12/03inbrief.html

（么媛媛编译，李红培 吴振新校对）

LC 格式可持续发展网站新加入地理空间格式描述

美国国会图书馆战略行动办公室宣布35个数字地理空间格式的描述和两个附属的短文已经开放获取。此信息是图书馆格式可持续发展网站的一个新部分。该网站提供数字内容格式的信息，强调与保存计划有关的一些方面和特征。目前该网站提供的信息有大约260种数字格式和子格式，分为7个种类：静态图像、声音、文字、动态图像、网络存档、数据集、地理空间。但仍有很多其它格式需要描述。

2010和2011年，从国家地理空间数字仓储（National Geospatial Digital Archive）开发的资料开始，Nancy Hoebelheinrich (Knowledge Motifs LLC) 和 Natalie Munn (Content Innovations LLC)开发了这个新的地理空间信息汇编。相关资料的链接如下：

http://www.digitalpreservation.gov/formats/content/gis_quality.shtml、

http://www.digitalpreservation.gov/formats/content/gis_intro.shtml

编译自：http://www.dlib.org/dlib/january12/01inbrief.html

（么媛媛编译，李红培 吴振新校对）

NEH 资助机会：数字保存和获取的研究与开发

NEH（National Endowment for the Humanities，国家人文学科捐赠基金）宣布要资助保存和获取方面的研究与开发（截止到5月16日），特别鼓励数字保存相关项目的申请。这个项目将提供总计$350,000的资金，利用三年时间来资助一些活动，比如开发一些技术标准、最佳实践、重复利用或者提高人文学数据获取水平的计算工具，以及保存人文学科馆藏的科学程序。在最近的资助周期内，特别关注三个领域：视听资源的保存和获取、预防性保存和数字保存。

考虑一下技术革新和新思想被引入和接受测试的快速程度，会发现数字保存或许能代表一切可能的活动。这个领域有很多的研究方向，一些最有趣也最复杂的研究领域包括但不限于：分布式或集中式网络、原生数字材料（比如社会媒体、当代艺术、数字报纸、数据库、数字人文学项目）的保存、数字资产和内容特别是视听材料管理系统的发展、文化遗产数据管理，以及能够增强人文学科馆藏交流和互操作的元数据标准的开发。

项目中“研究和开发”是指解决文化遗产团体中广泛存在的保存或获取问题，这或许是NEH项目中最“科学”的一个部分。NEH希望申请人能清楚地阐明该领域存在的问题和需求，制定一个包含一系列实验或者增量发展阶段的工作计划，然后尽可能地对研究结果进行

宣传。项目申请人要有一个长期可持续的项目计划，包括制定一个科学实验会在哪里实施的数据管理计划，或者制定通过附加项目阶段或与开源团体结合来开发一个可持续工具和标准的策略。只是单独地数字化大量馆藏或者生产相关资源的项目不能算作研究和开发。NEH在R&D计划中支持的项目主题必须与保存和（或）获取有关。

编译自：

http://blogs.loc.gov/digitalpreservation/2012/03/neh-funding-opportunity-research-and-development-in-digital-preservation-and-access/

（李红培编译，么媛媛 吴振新校对）

美国波士顿制作公司通过 Hydra 和 Fedora 将视听媒体变为未来的学术资源

任何使用应用程序的人都知道新闻和信息的记载、分享和传播方式经历了很大的变化。当发生一些不寻常的事时，人们去YouTube看看有没有音频或视频的第一手资料的可能性跟去看报纸的可能性几乎相同。制作高成本影片和即时视频来报道和分享世界上正在发生的事情，已成为每个拥有手机的人每时每刻（24/7）都可以参与的事情。蓬勃发展的原生数字媒体馆藏将来很可能会变成学术研究材料，让后代了解我们的日常生活文化和遗产。然而对资产不雄厚的文化机构来说，视频和音频资源的保存非常困难，他们缺少资金来对其进行妥善存储、转移、保存和提供访问。

美国波士顿制作公司(WGBH)最近得到了NEH保存和访问的研究与开发项目的资助，能够开发一个媒体保存的开源数字资产管理系统来帮助解决上述问题。如果图书馆保管员和档案管理员有合适的工具来管理和保存这些数字资料，那么诸如数字文件之类的媒体的可移动、可保存和可获取性会呈指数倍数增长。

这个项目将在Hydra的基础上，利用一个容易实施且易于与保存存储系统挂钩的Fedora仓储库，为媒体文件开发一个数字保存系统。这个堆栈有那些缺乏技术的公共媒体组织所需的灵活的数据管理、工作流模型和界面，最适合管理大型媒体文件。长期保存这些当前创造出的、将来会成为历史资料的原生数字媒体资料需要很复杂的工作流程。

WGBH采取这个解决方案的目的是建立一个适应公共媒体机构需要的能够广泛实施的系统，确保现在已创造的原生数字媒体资料能够被保存，在将来成为历史资料。

编译自：

http://duraspace.org/wgbh-turning-audio-and-visual-media-scholarly-resources-tomorrow-hy

dra-and-fedora

（李红培编译，么媛媛 吴振新校对）

510 家出版商参与全球 LOCKSS 网络

通过全球LOCKSS网络（Global LOCKSS Network，GLN），图书馆可以构建包含所有开放获取出版物及其订阅的电子期刊和书籍的集合并对其进行长期保存。GLN可以看作是全球图书馆资源的一个总集合。

已经有来自510家出版商的超过9000种电子期刊选择GLN作为它们的数字保存和长期获取的合作伙伴。相关合作出版商及其电子期刊出版物（标有ISSN 和eISSN）列表参见：

Download in comma-separated values format (.csv) (UTF-8 encoding)（更新于2012年2月22日）。

编译自：http://www.lockss.org/community/publishers-titles-gln/

（齐燕编译，李红培 吴振新校对）

合作：NDSA 的关键价值

2010年7月国家数字化管理联盟（National Digital Stewardship Alliance，NDSA）正式成立至今，已经有一年半多的时间了。从那时起，五十五个创始成员就致力于维护会员利益、共同制定组织目标和价值标准。如图1所示的五个工作组都已根据他们的共同利益制定了相应的工作计划，各个参与者们也都积极地执行各自的任务。详细信息参见：

http://blogs.loc.gov/digitalpreservation/2012/01/partly-cloudy-trends-in-distributed-and-remote-preservation-storage-more-results-from-the-ndsa-storage-survey/

http://blogs.loc.gov/digitalpreservation/2012/01/what-does-innovation-look-like-the-ndsa-innovation-working-group-wants-to-know/

http://blogs.loc.gov/digitalpreservation/2012/02/box-out-taking-digital-preservation-outreach-resources-to-the-classroom/

图1 NDSA 工作组角色图

现在，NDSA的会员已经达到110个，并且还在增加中，详细信息参见：http://www.digitalpreservation.gov/partners/。NDSA汇集了各方重视数字信息的长期保存及持久获取的从业者、专家以及利益相关者，它不仅仅是名义上的一个联盟。

这也是NDSA为什么在组织文件的制定中将“合作”确定为联盟的一个关键价值。随着联盟的扩大，许多成员自愿地腾出时间、分享知识，共同深入地讨论之前已经分享的价值观念，并制定出一套核心价值的陈述声明，供会员参照遵循。

作为这些努力的一部分，成员们会去审查其他志愿虚拟组织：看看他们的组织架构、入会资格、既有福利及其价值观念的具体呈现。很明显，作为一个新兴组织，没有一个现成的模型可以采用。但是却有一个内在的共同认知，那就是要将资源优化配置、共享专业知识、建立紧密团结的参与者团体作为价值观的核心。

NDSA是一个没有其它模型可以参照只能自己摸索发展的组织，但是它的会员会相互合作，相信最终将引领数字保存领域的发展。

编译自：

http://blogs.loc.gov/digitalpreservation/2012/02/collaboration-looking-across-the-ndsa-membership/

（齐燕编译，李红培 吴振新校对）

NDSA 存储调查结果：文件不变性和数字保存存储设备

数字对象的一个令人烦恼的特性就是难以确保其持久的真实性和稳定性。文件在使用过程中可能会被毁坏，即使不使用时也会出现比特腐烂现象，而传输过程中也可能会丢失对其

可操作性至关重要的部分。在最基本的层面上，数字资源长期保存要求我们相信我们现在正在使用的数字对象与之前使用的是一样的。

要解决这个问题，数字资源长期保存领域的人们会经常谈到数字对象的不变性。在这里，不变性是指始终如一的、稳固不变的、持久安定的性质。值得庆幸的是，在内容管理领域有很多用于数字对象检查的好方法，可以确保其内容质量。不变性检查是一个检验数字对象是否被改变或者毁坏的过程，在实践中，最常见的是通过计算和比较校验和或哈希值来实现。关于不变性检查方法的更多信息参见："Hashing out Digital Trust"。

NDSA 成员进行不变性检查的方法

国家数字管理联盟（NDSA）是一个致力于确保数字信息持久访问的合作伙伴网络。该联盟的目标是要建立、维护并不断提升保存国家数字资源的能力，以造福于当代及子孙后代。在过去的几个月中，NDSA 的基础设施工作组一直在报告对成员们长期保存存储状态调查的结果。之前有文章分别讨论过 "存取要求（access requirements）"及"云存储和分布式存储（cloud and distributed storage）" 在构建长期保存基础设施中的作用。调查中浮现的另外一个关键主题是，不变性检查作为性能需求的普遍性以及由此带给存储系统的挑战。

令人欣慰的是，基于了解正在保存对象的有效性和一致性的基本需求，有 88%的成员回应正在对其保存的内容进行若干形式的不变性检查，这表明，人们已经将其视为数字资源长期保存工作流程的重要组成部分。

同时需要指出的是，NDSA 的成员对其各自内容进行不变性检查的方法大相径庭。这些差异通常是出于种种较为复杂的原因，包括不变性检查软件的可扩展性、网络的限制和数据传输的成本、处理量和存储需求，以及与某些特定内容的可用性和管理相关的其它环境性因素。

在做出回应的受访者当中，不变性检查的实践情况如下所示（其中有些会员会采取多种策略）：

- 88%（49/56）的机构报告称，他们正在对其保存的内容进行若干形式的不变性检查。
- 57%（32/56）的机构会在某些处理（比如摄入）的前后进行不变性检查。
- 43%（24/56）的机构会周期性地进行不变性检查。
- 32%（18/56）的机构则采取随机抽样的方式检查其内容的不变性。

虽然不变性检查很普遍，但是 NDSA 成员们实施这些检查的时间安排各有不同，其中有 24 家机构拥有各自的固定计划：

- 46%（11/24）的机构至少每月进行一次内容不变性检查。

- 21%（5/24）的机构至少每个季度进行一次内容不变性检查。
- 29%（7/24）的机构是每年进行一次内容不变性检查。
- 4%（1/24）的机构是每三年进行一次内容不变性检查。

不变性的未来

NDSA 基础设施工作组成员多次提到，不变性检查的当前技术能够实现分布式的不变性检查，以及针对有意或无意的毁坏进行频繁的、强健的修复。这可以通过复制存储在多机构协同分布式网络中伙伴的分布式“镜像”存储库里的已校验的数据将毁坏的数据替换掉来实现。一些联盟组织如 MetaArchive 和 Data-PASS 使用 LOCKSS 来进行这种分布式不变性检查和修复。同时，一些个别机构也会使用自维护的分布式存储库系统来将毁坏的内容替换成经过校验的、未腐败的完好副本。

如前所述，这个 NDSA 工作组的主要兴趣之一是探究云存储系统在数字资源长期保存的存储架构中的潜在作用，对于那些使用中的云存储系统，遵照不变性要求对其进行评估时，会发现是有问题的。正如 David Rosenthal 指出的，目前的云服务无法证明它们不是简单地“重播”在存储之初创建和保存的不变性信息的，他强调：云服务商们需要提供一种工具或服务来供人们核查系统内存储着相关内容而不是仅仅缓存了不变性元数据。履行这种承诺是非常困难的，因为在各种云存储平台上运行任何一种频繁的内容不变性检查都将会付出相当昂贵的代价。

拥有像自动化的不变性检查和修复这种嵌入式功能是未来长期保存存储系统最为期待的特质。这个愿望的实现会遇到一系列挑战，比如人们需要处理系统类型的依赖性以及当前各种不变性检查流程上的差异问题，这体现了获取、性能、保存要求、存储基础设施等与机构资源之间复杂的相互作用。随着不变性检查变得无处不在，像分布式存储这类新选择会获得更多的认可，而相应的硬件支撑也会得到人们的呼吁以满足新的需求。

NDSA 希望浏览过上述这些决策的长期保存管理人员，在遇到相似的复杂性问题时，能够从其它 NDSA 成员的知识和经验中获益并能够设计出新的解决方案。

编译自：

http://blogs.loc.gov/digitalpreservation/2012/03/file-fixity-and-digital-preservation-storage-more-results-from-the-ndsa-storage-survey/#__utma=69962757.1872740993.1331287447.1331287447.1331290665.2&__utmb=69962757.10.9.1331291092417&__utmc=69962757&__utmx=-&__utmz=69962757.1331287447.1.1.utmcsr=(direct)|utmccn=(direct)|utmcmd=(none)&__utmv=-&__utmk=107330501

（齐燕编译，李红培 吴振新校对）

NDSA 设立创新奖

作为一个由多种成员群体组成的、共同恪守数字保存承诺的组织，NDSA 深知创新和冒险在推进和支持数字保存活动获得更大成功上的重要性，为此，NDSA 设立了“NDSA 创新奖”。

NDSA 已经确立了一些年度奖项来鉴定并鼓励数字保存管理领域的创新。这些奖项将会凸显并表彰为数字资源长期保存领域做出独创性或卓越贡献的富有创造力的个人、项目、组织和未来的管理者等。

这些奖项主要关注以下几个领域的卓越表现：

- 个人：为数字保存团体做出了显著的、具有创新性的贡献。
- 项目：其目标或产出展示了对数字保存管理工作更深入的、具有创新性的认知，或者为实现成功的、可持续的数字保存管理工作所必需的进程做出了突出的、革新性的贡献。
- 组织：采用一种革新性的方法为数字保存团体提供支持和帮助。
- 未来管理者：特指学生，但也包括教育工作者、培训师、课程开发人员等，采用了一种创新性的方法来推进与数字保存问题和实践相关的知识的发展。

NDSA 认识到数字保存管理的创新会有多种形式，所以这些奖项的参选资格设置得非常宽泛。NDSA 希望借此发现并奖励那些应对数字保存挑战的新奇的、冒险性的、革新的方法。

编译自：

http://blogs.loc.gov/digitalpreservation/2012/03/announcing-the-ndsa-innovation-awards/

（齐燕编译，李红培 吴振新校对）

法国的数字资源长期保存

多年来，法国数字资源长期保存领域的从业者们一直努力营造和维持一种社区归属感，同时，相关的国家政策也不断出台。人们可以在诸如国家兴趣小组季度会议上交流各种消息和最佳实践。

法国数字资源长期保存领域的“蓝图”已经确定下来，主要分为三个部分。一些存储机构已率先开发了大型数字资源仓储库。受到 CCSDS 构建的 OAIS 模型的启发，国家空间研究中心（CNES）成为先锋者之一。法国国家图书馆 (BnF, http://www.bnf.fr/en/professionals/preservation_spar/s.preservation_SPAR_presentation.html) 和国家高等教育 IT 中心(CINES,http://www.cines.fr/spip.php?rubrique219) 也纷纷效仿，并且更关注资源的关联化。CINES 已经运行了一个存储数字博士论文的国家级平台，并掌控着多所大学的数字化馆藏。BnF 的初衷是要将它每年数字化的大量馆藏资源摄入到其资源仓储库中，不仅是自身持有的还包括其合作伙伴的资源（部分可获于：http://gallica.bnf.fr/）；今年，它将开始其网络存档资源的长期保存。但是，它也开发了作为第三方存储机构的能力，并已经有一些其他存储机构成为了其首批客户。另外，国家视听研究院 (INA, http://www.ina-sup.com/en) 保存了法国广播和电视节目的所有数字副本，同时存储了与各家电台和电视运营商相关的网络内容。

政府机构，特别是那些负有长期存档责任的机构，如国家或区域档案馆，或一些负责处理司法证据的组织等，也赶上了数字资源长期保存的步伐。自从 2000 年有关推行无纸化办公的法律颁布以后，它们不得不开始进行数字资源的长期保存。这些工作已经催生了一系列标准，比如存档过程中的数据交换标准(Standard for Data Exchange in Archiving，SEDA)，用来规范记录的转移、修改和删减，这一标准的实施是由 XML schemas 实现的。

近期，越来越多的解决方案正在被开发出来以满足私营企业的需求。传统的存档公司正在扩张其业务技能来应对其客户的数字需求，如软件编辑公司 Naoned，现在它的产品能够同时处理纸质和数字化的记录。相反，一些规模或大或小的 IT 咨询公司，也已经将数字存档列入其业务范围中，比如 ATOS，它利用从开发 BnF 的资源仓储库中获取的经验来开展自己的长期保存业务。这些公司与那些已经建立了记录存档标准的公共合作伙伴一起参与了标准化的联合活动，使之逐渐被国际标准化组织所采纳，即 ISO14641-1。同时，这些公司也在从事定义数字安全的特征的工作，以满足客户的需要，尤其是那些特别关注自身记录可靠性的客户。资源仓储库的认证工作也在进行之中。

越来越多的工作仓储库正在投入使用，同时，数字资源长期保存也逐渐被列入一些培训课程的教学大纲。现在应该正是一个参与国家和国际长期保存社区并互相分享经验的好时机。

编译自：

http://www.dpconline.org/newsroom/whats-new/798-whats-new-issue-42-february-2012

（齐燕编译，李红培 吴振新校对）

英国研究型图书馆与保存咨询中心的联合研讨会

英国研究型图书馆（Research Libraries UK）和保存咨询中心（Preservation Advisory Centre）联合举办了一次研讨会，来探讨研究型图书馆的资源保存保护方法与其馆藏管理的全局战略间的关系。为期一天的研讨会审查了研究型图书馆是如何处理决策制定、风险评估和优先次序等馆藏问题。参会者讨论的主题涉及将保存作为当前和未来战略之一、使用数字代理作为保存策略，以及与保存国家藏品的需求和安排相关的问题等，主要包括：

1、 专题陈述

1) 当前，在何种情况下你可以使用数字代理作为一种保存策略？（http://www.bl.uk/blpac/pdf/safegreen.pdf）

2) 伦敦大学国王学院的资源保存的优先次序（http://www.bl.uk/blpac/pdf/safesambrook.pdf）

3) 将保存纳入战略规划中（http://www.bl.uk/blpac/pdf/safecheckleyscott.pdf）

4) 当你丢弃某些东西时，你都做了什么？是否做出了错误的决定？（http://www.bl.uk/blpac/pdf/safefowler.pdf）

5) 明智的决定——Copac 馆藏管理工具项目（http://www.bl.uk/blpac/pdf/safeemly.pdf）

6) 保存国家藏品（http://www.bl.uk/blpac/pdf/safebanks.pdf）

2、相关资料

1) 是否处于安全控制之下？RLUK 实施联合保存所面临的挑战（http://www.bl.uk/blpac/pdf/safechallenges.pdf）

这篇文档考虑了过去和现在的情形，进而预测了由 RLUK 和英国图书馆保存咨询中心（British Library Preservation Advisory Centre）联合举办的为期三年的培训项目成功实施后的未来情形。RLUK 将会根据其当前战略规划对这篇文档、这次研讨会，以及联合项目的成果进行评估。

2) 保存生态环境——关系映射图（http://www.bl.uk/blpac/pdf/safemap.pdf）

保存政策是机构藏品保护战略的核心。它必须与机构使命及其它关键战略，比如藏品管理等相一致，必要时要相互参照，它也会体现在机构资源配置方法上。同时，在不同的子领域或专门领域中会有相应的下级政策和作业程序/协议对保存政策提供支持。这张关系映射

图总结了藏品管理实施过程中开展保存工作的相关因素，并包含了对其它一些信息资源的链接，来为制定了机构藏品保护战略下的保存工作提供相关支持。

3) 制定保存政策（http://www.bl.uk/blpac/pdf/safetemplate.pdf） 此模板可以为那些想要为其自身机构制定保存政策的员工提供支持和帮助。不同的组织机构，其具体需求不同，保存政策也会有所差异。政策制定人员在考虑要包含哪些方面时，也会特别希望能够参考保存咨询中心的保存政策构建模块指南，以及其它机构的一些案例。

编译自：http://www.bl.uk/blpac/safehands.html

（齐燕编译，李红培吴振新校对）

【重要文献摘译】

网络存档概述

Jinfang Niu 著 么媛媛 编译

摘要

该概述研究了多个大学、国际政府图书馆和档案馆为其存档选择、获取、描述和访问网络资源时所采用的方法。创建一个网络存档不是易事，图书馆和信息学校应该确保将网络存档方法和技巧列入其课程的一部分，帮助未来的从业者迎接那些挑战。为开设网络存档课程做准备，笔者进行了一次全面的文献回顾。该文章汇报了笔者的研究结果和其对一些正在使用的方法的看法，例如传统的存档管理概念和理论如何应用到存档的网络资源的组织和描述上。

介绍

网络存档是对记录在万维网上的数据进行整理、存储，确保这些数据保存在存档里的过程。互联网存档（Internet Archive，IA）和几个国家图书馆于1996年发起了网络存档的实践活动。于2001年开始举办的国家网络存档研讨会（International Web Archiving Workshop，IWAW）为分享经验和交换意见提供了一个平台。之后于2003年建立的国际互联网保存联盟（International Internet Preservation Consortium，IIPC）极大地促进了各国合作，并为创建网络存档开发标准和开源工具。这些发展以及人类文化在网络上被创建和保存的比例越来越

大，使得越来越多的图书馆和档案馆不可避免地要面临着网络存档的挑战。

2010年秋天，一项针对美国排名前32的图书馆和信息学校的课程目录的调查发现，只有一所学校——密歇根大学——开设了一个为期半学期的网络存档课程。印第安纳大学在其“网络内容分析”课程中涉及到网络存档这一主题。加州大学洛杉矶分校在其“数字记录管理”课程中涉及这一主题。虽然很多诸如美国伊利诺伊大学之类的学校有开设数字保存、数字化保管、Web2.0对存档理论和实践的影响之类的课程，但却不清楚它们能涉及到多少网络存档的内容。笔者认为，网络存档需要足够的专门知识和技能，所以需要开设一门单独的课程。

像许多其他种类的信息资源的管理一样，网络存档的工作流程包括评价和选择、采集、组织和存储、描述和获取。这一工作流程是网络存档的核心。虽然数字保存肯定是网络存档的整个流程中的重要一步，但它不是网络存档所独有的。网络资源的保存跟其他数字资源的保存没有什么不同，该文章中的调研不包括数字保存。

评估和选择

档案社区中的“评估”指的是评价记录的价值并决定记录是否应该被保存和应该被保存多久的流程。它实质上是一个选择的过程。该文中，评估被用作选择的同义词。所有的网络存档都基于一个或多个标准来选择要保存的网络资源。虽然 IA 想要保存整个网络，但实际上只是抓取了网络表面的网页。网站层次结构中比较下层的网页通常不会被 IA 抓取。

现在的网络存档工作采用以下的选择标准来决定该保存哪些资源：域名（如.gov或.edu）、主题或事件、媒介类型和流派。许多欧洲国家都在国家域名范围内保存网页。美国宇航局戈达德太空飞行中心（NASA Goddard Space Flight Center，NASA GSFC）抓取戈达德域的网页。美国国会图书馆已创建了多个基于时间的网页集合，如2001年9月11日的网络存档、2003年总统选举的网络存档和伊拉克战争的网络存档。基于媒介类型的选择指包括或不包括某些类型。例如，戈达德图书馆就不保存大型视频文件和软件产品。另一方面，Chirag Shah 和 Gary Marchionini 领导的网络存档项目侧重于保存 Youtube 上的选举视频。一些网络存档的选择基于类型（genre），如博客、报纸、虚拟世界等。法国国家图书馆创建了一个电子日记的网站收集。IA 有一个软件存档和一个视频游戏存档。保存虚拟世界项目特别对在线虚拟世界存档进行了研究。Antonescu 指出了保存在线虚拟世界的两个不同方法。一个方法保存了技术基础设施——虚拟世界中存在的对象和虚拟实体——而另一个方法保存了虚拟实体的互动和生活经验。Winget 和 Murray 进行了一项保存开发视频游戏过程中创造的记录和人工产品的研究。

理论上来说，基于客观标准的自动化选择是很容易的。在技术层面上，软件很好判别网络资源的媒介类型（音频、视频或文本）和域名（.gov 或.au）。同样地，区分在线杂志或博客之类的类型或判断博客文章和评论之间的差异也应该不难。高品质或流行的网络内容已经经过了无数的接入链接和访客、在线视频观众和用户评分的鉴定。捷克共和国国家图书馆将域名鉴定流程进行了自动化。

然而，基于主题或事件的选择就需要人工判断了。信息专家的手工选择既费时又昂贵，因此只适用在小型网络存档中。为了减少人工选择的成本，一些网络存档接受用户推荐的URL，利用现有的网址登记或邀请学科专家来帮助选择要保存的网络资源。澳大利亚保存和访问联网文献资源（Preserving and Accessing Networked Documentary Resources of Australia，PANDORA）和国立台湾大学图书馆档案馆接受用户推荐的网站。中国研究数字存档（Digital Archive for Chinese studies，DACHS）邀请中国的专家学者来评价相关网站。英国政府网络存档项目用所有英国中央政府网站的一个注册表来选择网站；注册表中的网址由网站管理者来提交和管理。

加快手工选择的另一个方法是利用存档管理领域的宏观评价理论。就像保存政府网站出版物的亚利桑那州模型中解释的那样，宏观评估必须基于集群的网页而不是单个网页来评价和选择网络资源。评估一个集群缩小了问题的规模，使评估流程效率更高。这种集群可以在不同层面上来决定。美国国家档案和记录管理局（U.S. National Archives and Records Administration，NARA）利用其对政府机构的指导中的几个分析单元来对网络记录进行风险评估：网站组、整个网站、减去显示出明显不同特点的一两个部分的单个网站、网页的聚类。这几个分析单元也能用在网络存档的选择中。例如，图书馆员和档案员可以评估整个网站的价值而不是单个网页的价值来决定这个网站是否应该被保存。

诸如域名和媒介类型之类的选择标准可以加入一个基于价值的选择或一个代表性抽样方法。国立台湾大学的网络存档收集了历史、文化、社会、教育或学术观点等方面非常有价值的网络资源。垃圾邮件过滤也是一种基于价值的选择方法。另一方面，代表性抽样避免了基于价值的选择的主观性和偏见，试图去建立一个该保存什么资源的代表性模板。Lyle 将抽样策略运用到爬虫下载的网络资源中，以减少其需要保存的数量。法国国家图书馆在抓取信息之前运用抽样策略来制定种子名单和过滤条件；该图书馆认为馆藏应该“反映多样性的法国社会和文化，不论有没有科学价值或是否流行”。因此，“网络存档应包括‘最佳’（文学、科学出版物）和‘最差’（从广告到色情）。小型、中型和大型资源被收集的机会应该是相同的。”

获取

根据网络存档的规模、网络存档和网站用户之间的关系和存档的网络内容的性质的不同，要用不同的获取方法。图书馆和档案馆有从政府机构、捐助者和出版商的合法存储中获取转让资源的历史传统。这个方法对网络存档同样适用。Adrian Brown 指出，数据库驱动的动态网站不适合直接转让，因为数据库通常是专有的并且难以长期保存。一个简单的方法是利用如 DeepArc 之类的工具将数据从专有数据库格式转换成诸如 XML 的开放标准格式。

爬行抓取是网络存档所运用的一种独特方法，靠爬虫从网络服务器中获取内容。爬虫用种子清单来下载网络内容，然后跟随着超链接来发现和下载更多的网络内容。选择决策是编制种子名单和配置爬虫参数的基础，就像是一个过滤器，只保存通过过滤之后的链接。在一些图书馆和档案馆获取网络出版物的过程中，爬行正在取代存储。因为爬虫的局限性，一些网络资源需要手动获取。例如，一些爬虫无法获取 GIS 文件、动态网络内容或流媒体。NARA 对不能用爬虫获取的特定网络内容记录格式所能采用的合适获取方式提供了一份指南。

重复爬行一些未更新的网页会造成网络存档的重复，浪费用来管理、存储和保存的物力。幸运的是，如今诸如 Heritrix 的最新版这样的智能爬虫可以在下载和存储网络资源时减少重复。重复爬行大型的、经常更新的网站会造成时间的不连续。爬行一个大型网站可能会需要几天甚至更长的时间，在此期间，网站还可能会更新。假设有一个网站有两个网页，爬虫在 t_1时间点下载了 p1网页，当爬虫到达 p2网页处时，p2和 p1分别更新成了 p2-a 和 p1-a。在这种情况下，原来的网站包括 p1和 p2，更新的网站包括 p1-a 和 p2-a，然而存档网页却是 p1 和 p2-a。也就是说，爬虫获取了一个从没存在过的网站。

获取网络资源时，决定是否要获取版权人的许可取决于网络存档的法律环境、规模、存档内容的性质和保存组织。像新西兰这样在法定存储中包括网络资源的国家，其法定存储图书馆在进行存档时不需要获取其国内创作的网络出版物的许可。像 NARA 和英国国家档案馆这样有法律授权保存公共记录的政府档案馆也不用从记录创造者那里获取许可。在相同的法律环境中，获取小规模网络存档的许可的可能性大于大规模的网络存档，因为所有权人的数量相对较少。诸如 IA 这样的大规模网络存档倾向于应用退出机制，Hal Varian 认为，谷歌图书馆项目的退出机制是很明智的，因为获取许可时的选择模型中的交易成本太高了。这种观点同样甚至更适用于网络存档，因为多数网络存档并没有从保存网络内容中获取经济利益，而且难以确定匿名创建的网络内容的所有权人。

组织和存储

网络存档需要保存网页内容的真实性和完整性，这一需求根据收集目的的不同而变化。

在某些情况下，只保存知识内容就够了，在其他情况下，像资源的法律证据、结构和背景这样的内容也需要保存。根据传统的存档管理理论，档案记录的背景包括出处和原始顺序。出处包括记录的源信息，如记录生产者、导致记录发生的交易、产销监管链等。原始顺序是记录生产者或记录管理者最开始对记录之间关系的顺序安排。

对于网络资源，原始顺序的概念可以与传统存档管理理论中结构的概念相结合。原始顺序基本上就是存档的网络对象的外部结构。传统的存档管理理论中定义的结构基本上可以认为是存档的网络对象的内部结构。如，对一个存档的网站来说，它的外部结构显示了该网站是如何关联到其他网站的，也就是网站的原始顺序。从外部网站来的链入链接和该网站与其他网站的链出链接是这个外部结构的一部分，因此是该网站的原始顺序。该网站的内部层次结构显示了其组件和子组件是如何关联的，这就是传统存档管理理论中定义的结构。内部结构是用网站内部的超链接来定义的。对于一个如网页这样存档层次较低的对象来说，外部结构显示的是它怎样与其他网页相关，内部结构显示的是该网页的内部组件，如文本内容、图像、音频、视频等是怎样安排的。在重复获取中，也有显示着网络内容演变的历史背景，包括网页的新旧版本。

Masanes 认为，当前的网络存档主要采用三种方法来组织和存储存档的网络内容：本地文件系统、基于网络的档案和不基于网络的档案。这三种方法都保存了网页的知识内容，但根据不同的背景和结构，保存的程度不同。

在使用本地文件系统的网络存档中，浏览器可以像浏览网页一样浏览文件系统。除了那些在网络存档的范围之外的未存档的链接，网站的内部层次结构和不同网站之间的链接关系都被保存了下来。然而，为了让网络资源和文件系统相互适应，需要做两次背景转换。首先要修改 URI 的命名规则使其符合本地文件系统的要求，然后，要将绝对链接转化为相对链接来允许在文件系统中进行浏览，否则绝对链接将指向现有的网页而不是保存的网页。

在基于网络的存档中，网页和相关元数据被分组储存在文件中，原始 URI 和链接也被保存起来了。虽然链接还需要重定向或转化为指向存档网络而不是现有网络，但只在用户访问这些链接而不是在存档中写入这些链接时才需要这么做。这种方法最大程度地保存了真实性。

不基于网络的存档是从超文本背景中提取网络文件并将其重新组织成基于目录的访问模式或将其转换成 PDF 文件的方法。该方法保存真实性和完整性的程度是最低的。

描述和元数据

元数据生成方法和丰富的元数据生成取决于网络存档的规模和保存组织现有的资源水

平。非常大的网络存档通常依赖于自动生成元数据。诸如获取网络资源时产生的时间戳、状态代码、尺寸大小、URI 或 MIME 类型的元数据信息都可以用爬虫来获取或创建。也可以从 HTML 页面的元标签中提取元数据，即使由于搜索引擎优化选择的关系一些元标签可能并不准确。希腊网络存档项目从网页和锚文本中自动提取关键词，然后用这些关键词来将网页分成不同的集群。

小规模的网络存档可以手动创建元数据。加州大学洛杉矶分校的在线竞选文学存档使用都柏林核心元数据、美国国会图书馆的主题词和本地定义的权威名单。其管理元数据来自于工作人员抓取和审查程序时创造的详细记录。中国研究数据档案馆网络存档邀请学者贡献一些描述性元数据。国立台湾大学网络存档特意为网络内容创建了一套三层次的分层分类方案和编目规则。元数据也可以通过用户标记、注释或评级来创建。美国国会图书馆基于 URL 推荐所创造的元数据自动生成元数据对象描述模式（Metadata Object Description Schema，MODS）的记录，然后通过编目来改善这些记录。

网络存档集的结构是多层次的。一个网络存档集可能包括好几个爬行线程（session），每个爬行线程包括几个网站，每个网站又包括几个网页，每个网页还可能由许多如文本文件、图像文件和视频文件之类的文件组成。这种层次结构跟存档集的层次结构相互匹配，存档的多层次描述方法可以应用到保存的网站中去。存档社区使用自上而下的方法：首先为较高级别创建元数据，然后如果资源是可获取的，就为较低级别创建元数据；为较低级别创建的元数据可以继承较高级别的；基本不为条目（item）级别的对象创建元数据。该自上而下的方法和元数据继承机制也适用于网络存档。此外，一些条目级别对象的元数据，如文件格式、尺寸大小和修改日期等，也都可以自动提取。

网络存档决定使用书目方法，只创建单一层次的描述时，应基于网络存档和可用资源的规模选择描述单位。选择像整个网站这样较高层次的描述单位会减少描述细节和创建的元数据记录。美国国会图书馆和哈佛大学网络存档给包含很多网站的网络存档集创建了一份 MARC 记录，该记录可以通过图书馆目录来检索。选择像页面级别这样较低层次的描述单位会增加描述细节和创建的元数据记录。除了 MARC 记录以外，美国国会图书馆网络存档还为网站创建了 MODS 记录。这些 MODS 记录可以在网络存档中进行检索，但不能通过图书馆目录访问。PANDORA 也选择一个网站和网站的一部分作为描述单元。

访问和使用

对网络存档的访问取决于存档所在国的国家法律环境。新西兰的合法存储法允许新西兰国家图书馆保存任何新西兰网站公开提供的页面并提供该存档的副本。在美国，美国国会图

书馆给所有可公开访问的存档网站做了书目记录，并且只允许公开访问生产者已许可访问的网页。许多网络存档都是死档或者只能到馆访问，如法国国家图书馆网络存档和法国视听国家研究所（Institut National de l'Audiovisuel of France，INA）、芬兰网络存档、丹麦网络存档（Netarchive.dk）、挪威网络存档、斯洛文尼亚网络存档、瑞士网络存档和奥地利网络存档。一些可公开访问的网络存档通过减少功能或延迟访问来避免与网站所有者的冲突。

不同的网络存档的搜索能力取决于其元数据的丰富程度和其使用的搜索与索引工具。美国国会图书馆网络存档和新西兰网络存档支持通过认证控制的访问界面来进行检索，因为实际上这两个存档都用了主题词作为元数据记录。另一方面，基于 Wayback Machine 的网络存档只能通过 URL 来搜索，而基于 NutchWax 搜索引擎的网络存档也支持全文检索。已有人创建了一些高级访问界面。英国网络存档通过挖掘内容、标签云和3D 墙创建了两个可视化界面。Jatowt 等人也尝试了几个能显示网页的历史版本的高级方法，他们创建了一个幻灯片和一个二维图形来显示 URL 随时间而演变的不同内容。

结论和下一步计划

现有的网络存档有不同的选择、获取、组织、存储、描述和提供访问的方法。这种差异一方面是由外部因素造成的，如法律环境、网络资源生产者和网络存档之间的关系等；另一方面还有内部因素，如存档的网络内容的性质、存档组织的性质、网络存档的规模、存档组织的技术和经济能力等。

本概述是基于对介绍网络存档如何实现的文章的全面研究而写成的。然而，没有哪篇文章直接解释了该领域进行日常选择、获取和编目网络存档工作的专业人员所需的知识和技能。笔者正在策划一个研究项目来填补这一空白，将采访一些正在做这些事情的图书馆员和档案员。相关从业人员的观点将为网络存档课程的设计提供更多有价值的信息。

编译自：http://www.dlib.org/dlib/march12/niu/03niu1.html

（李红培 吴振新校对）

面向研究人员的网络存档：呈现方式、期望和潜在用途

Peter Stirling，Philippe Chevallier，Gildas Illien 著 么媛媛 编译

从 2006 年起，互联网就已经被纳入到法国合法存储的法规中，网络存档成了法国国家图书馆（Bibliothèque nationale de France，BnF）的任务之一。2008 年起，该图书馆已经在

实验的基础上提供对网络存档的访问。许多国家对网络存档的兴趣越来越浓，尤其是关心它怎样能最好地为研究人员服务。基于 BnF 对网络存档的潜在用户，特别是工作在与互联网相关领域的不同研究者进行的访谈，研究人员曾在社会科学的互联网研究这一扩展领域中对网络存档做了一个定性研究。该研究旨在探讨他们在内容和服务方面的需求，还分析了这些存档的不同呈现方式，以确定提高存档利用率的不同方式。虽然对研究人员来说，维持网络“记忆”的兴趣是显而易见的，但他们面临着界定什么是有意义的文件集合的问题，这个范围似乎是无限大的。诸如国家图书馆之类的文化遗产机构可以作为可信任的第三方来创建构造合理且存档完好的集合，但是这些存档有一定的道德和方法问题。

虽然这项研究是基于有限的采访对象的，但他们的反馈所提供的很多元素都有利于 BnF 规划其互联网法定存储的未来发展。

文章主要内容包括：

- 研究人员对于 Web 的使用；
- 研究人员对于网络存档的看法；
- 网络存档的内容与选择；
- 研究人员的潜在服务和信息需求；

文章的最后，作者进行了总结，提出了期望：

关于内容和选择政策，研究人员们都认为无法预测将来什么资料会吸引专业的或业余的研究者。但是如果已知现存数据的数量的话，做出某种程度的选择也是合理的。在这方面，研究表明 BnF 所采用的选择方法——将大规模抓取和基于手动选择的重点抓取结合在一起的“混合模式”——似乎是应对这种矛盾的最好选择。应该维持这项政策，然而，抓取必须更能对网络的演变做出反应：追踪节点和网络、最受欢迎的站点、Google 搜索结果等。最重要的是既要保存单个元素，也要保存网络上显示新趋势的社会或商业活动。

关于服务和推广，选择政策中所采用的决策和标准都必须是合理的、有证明的和可见的。要有工具能使研究人员了解一个站点是否已被保存并能找到和区分各个存档。由于网络存档对大多数研究人员来说都是一个新概念，所以需要探索其不同的比喻义来让不同的用户群体描绘和构想网络存档，以便加强其交流和推广。

最后，团体和合作是至关重要的。这不仅包括研究人员通过参与会议或其他机构和项目来进行交流，也要将研究人员和从事于网络相关工作的业余人员结合在一起，让他们参与到存档的方法和来源界定的创造中去。研究团体不仅要提高网络存档在科研工作中的“合法性”，也要负责鼓励研究人员有效利用网络存档中的资源。

全文请见：http://www.dlib.org/dlib/march12/stirling/03stirling.html

（李红培 吴振新校对）

网络存档的功能

Jinfang Niu 著 么媛媛 编译

网络存档有些功能对用户很重要，从基础的搜索和浏览到高级的个性化和定制服务、数据挖掘和网站重建。本文作者研究了截至2011年4月的十个最成熟的英文网络存档，以确定每个存档支持的功能并进行横向比较。基于 IIPC（International Internet Preservation Consortium，国际互联网保存联盟）的用户案例和两个相关的用户研究，笔者设计了一个功能清单。随后进行了一次功能审查，还对网络存档方法进行了全面的文献回顾。该文章描述了清单中所使用的功能以及这些功能被各个不同的网络存档运用到了何种程度，并讨论了笔者的调查结果。

功能简介

搜索参数

- **通过 URL 和关键字搜索**
- **基于域的搜索**
- **按日期缩小搜索范围**
- **按媒介类型缩小搜索范围**
- **一体化搜索**
- **不支持设定搜索参数的**
- **搜索界面的可用性问题**

搜索结果

- **按域对搜索结果进行分组**
- **持久性标识**
- **显示存档内容**
- **打印**
- **可靠性认证**

浏览功能

政策相关功能

个性化服务

数据挖掘

遗失网站重建

在研究中发现但未列入功能清单的功能

- **副本管理**
- **通过搜索引擎来发现**
- **显示未存档内容**

各个网络存档通常都支持一些基本功能，如按 URL 和关键字搜索，按日期、域、媒介类型缩小搜索范围等。然而，调查的这些网络存档都不支持一些高级功能，如数据挖掘、个性化服务、遗失网站重建等。有几个网络存档有明显的可用性问题，如看不到帮助链接、隐藏了高级搜索工具和分段搜索等。

调查结果还显示了网络存档的功能的适用范围及其与 IIPC 使用情况和用户的确定期望的比较。这些问题的存在并不一定意味着那些网络存档不关心或没有能力解决它们，更可能的是他们的注意力正集中在馆藏建设上，没时间提供更高级的功能和改善其网络存档的可用性。尤其是对于那些新的网络存档来说更是如此。

作者认为，随着时间的推移，存档的功能和可用性会得到改善。事实上，当时隔四个月，笔者2011年8月重新访问这些存档时，发现了一些已经改善了的地方。例如，英国网络存档增加了一项 N 元语法搜索功能，可以给出一个显示很长一段时间里英国网络存档中出现过的搜索项的图表。英国政府网络存档也增加了一项自动通过主题对检索结果进行聚类，让用户可以按主题过滤检索结果的功能。

编译自：http://www.dlib.org/dlib/march12/niu/03niu2.html

（李红培 吴振新校对）

【技术与工具】

FIDO 1.0.0 版发布

FIDO（Format Identification for Digital Objects）是一款 Python 命令行工具，能够高性能地识别数字对象的格式，并易于集成到自动化工作流程中。

它将对象特征转换为正则表达式并直接应用。FIDO 是免费的，易于安装，可以在

Windows 和 Linux 上运行。最重要的一点是，它的处理速度非常快。其核心是存储在三个文件中的几百条代码，运行时占用的内存非常小，在 XP 系统下，无论是识别 5 个文件还是 5000 个文件，它的内存消耗不足 5MB。

新的版本在编码和功能上进行了较大改进，详细信息参见：

https://github.com/openplanets/fido/zipball/master

http://www.openplanetsfoundation.org/blogs/2012-02-27-fido-version-100-released

编译自：

http://www.openplanetsfoundation.org/blogs/2010-11-03-fido-%E2%80%93-high-performance-format-identifier-digital-objects

http://www.dpconline.org/newsroom/whats-new/814-whats-new-issue-43-march-2012#whatsnew43

（齐燕编译，么媛媛 吴振新校对）

小型机构数字内容的分布式存储

据全国各地的数字保存推广和教育(Digital Preservation Outreach and Education，DPOE)项目组工作人员反映，数字保存工作进展缓慢，尤其对于一些小型机构来说，履行其数字内容长期保存的责任更具挑战性。在一次培训活动上，针对“无法运行 LOCKSS 的小型机构如何开展存储和保存工作”这一议题，有人推荐了将数字内容进行分布式存储的三种途径，包括：

1) 利用流行的联机文件存储服务，方便用户间的文件共享。在这种模式下，文件被存储在服务提供方，当然用户也可以选择存储在本地硬盘上。
2) 将内容拷贝到硬盘上，然后将硬盘邮寄到异地合作伙伴那里来妥善保管。
3) 利用虚拟主机服务存储内容，这时需要特别注意附属细则，因为这些服务通常既不能保证所存储文件的保密性也不能保证其安全性。

不可否认，最后一种方法不适用于小型机构，但是它也体现了利益来自合作这一主题。小型机构可以通过与各种图书馆服务部门进行合作，联合各类资源来创建一个机构知识库。

更多相关资料可以加入 DPOE 邮件列表获取。

编译自：

http://blogs.loc.gov/digitalpreservation/2012/03/talking-about-storage-solutions-for-the-%e2%80%9clone-arrangers%e2%80%9d/#__utma=69962757.2088657173.1333090678.1333090678.

1333094415.2&__utmb=69962757.4.9.1333094783849&__utmc=69962757&__utmx=-&__utmz=69962757.1333090678.1.1.utmcsr=(direct)|utmccn=(direct)|utmcmd=(none)&__utmv=-&__utmk=108049477

（齐燕编译，么媛媛 吴振新校对）

【资料推荐】

DPC 发布最新技术观察报告：《长期保存电子邮件》

电子邮件是数字时代标志之一，它的出现彻底改变了人们沟通交流的方式，是具有里程碑意义的一项信息技术。现在，企业依靠电子邮件开展业务、家庭和朋友通过电子邮件加深感情、政府部门也需要电子邮件支持无纸化办公……可以说没有其它技术能够像电子邮件这样普及。电子邮件内容承载了海量的私人和正式的交流信息，其中有些是“珠宝”，当然更多的也是需要丢弃的“废品”，人们越来越注重如何对其进行有效的管理。

机构、组织和个人都需要相当多的投入来确保电子邮件的安全。IT 管理者和档案保管员们很早就认识到，要想确保电子邮件的长期可用性，就需要对其进行周密的管理。然而，有关具体实践操作的实用性建议却相当匮乏，为此，DPC 精心制作了其最新的技术观察报告——《长期保存电子邮件》。

该报告全面地介绍了所有需要长期保存大量电子邮件的机构或个人需要注意的问题。伦敦大学国王学院的加雷思 奈特（Gareth Knight）认为《长期保存电子邮件》是一份极佳的主题综述，它将多个研究项目的观察结果综合在一起，对确保电子邮件数字资产长期保管和保存工作的顺利开展所必须解决的法律、技术和文化问题进行了简明的综述。它的结论强调了电子邮件保存的复杂性，为个人和机构更好地管理他们的电子邮件存档提供了很多实用的、通俗易懂的建议。同时它也指出：这是件人人都可以做到的事情！

大英图书馆是最近正在致力于制定新的电子邮件保存策略的机构之一。莫琳·彭诺克（Maureen Pennock）认为该报告中牛津大学博德利图书馆和医学研究委员会的两个案例研究非常有价值，能够帮助保存单位更好地理解其面临着的实际问题，以及如何切实有效地解决它们。这两个案例展示了长期保存电子邮件可以达到的效果，并强调了电子邮件标准的重要性。”

技术观察系列报告的总编辑和主要研究员、Charles Beagrie 咨询有限公司的 Neil Beagrie 指出，今后会陆续发布更多的技术观察报告，计划出版 5 个技术观察报告，全部出自顶级专家，《长期保存电子邮件》是第一份，今年还会出版《移动影像长期保存》、《数字保存知识产权》、《数字取证和保存》、《电子期刊的保存信任和持续获取》等。报告版式有了变化，添加了 ISSN 和 DOI 标识，以提高其使用率、引用率及影响力。DPC 和 Charles Beagrie 都希望这些报告能够为全世界的数字保存和最佳实践的普及做出巨大贡献。”

牛津大学博德利图书馆副主任兼 DPC 主席认为今年是 DPC 十周年纪念，作为标志性的事件之一就是发布一些新的报告来更新并扩大联盟所提建议的影响范围。技术观察报告可以为所有对确保其数字藏品长期可获取感兴趣的机构和个人提供实用的、持续的帮助。DPC 在保证其可获取的同时也会确保报告内容的权威性。电子邮件的长期保存几乎与每个网民都有关系，所以这是 DPC 十周年纪念的一个伟大的开端。

该报告 URL：http://dx.doi.org/10.7207/twr11-01

报告目录：

编译自：

http://www.dpconline.org/newsroom/latest-news/805-email-tomorrow-and-next-year-and-forever-preserving-email-report-published

（齐燕编译，么媛媛 吴振新校对）

可持续性状态：NDIIPP（2012）赞助州立项目述评

对“美国国家数字信息基础设施和保存计划 (National Digital Information Infrastructure andPreservation Program，NDIIPP) ”赞助的多个州立项目进行评审，包括分析项目产出和相关文档、在会议和专业性活动上与项目参与者进行个别交流、访问主要合作伙伴的网站，以及实时监测项目活动和公告等。北卡罗来纳大学教堂山分校的 Christopher A. Lee 总结了评审过程中的几点发现，形成了此份报告。

报告 URL：

http://www.digitalpreservation.gov/multimedia/documents/ndiipp-states-report032612_final.pdf

报告目录：

编译自：http://www.digitalpreservation.gov/multimedia/publications.html

http://www.digitalpreservation.gov/multimedia/documents/ndiipp-states-report032612_final.pdf

（齐燕编译，么媛媛 吴振新校对）

Duraspace 提供“了解未来：数字保存规划”系列活动资料

2012年2月-3月，举办了三场题为“了解未来：数字保存规划”的热门话题系列网络研讨会。

第一场该研讨会中，发言人主要关注保存规划的流程和最佳实践。

在线观看请访问：

http://www.slideshare.net/DuraSpace/assessing-preservation-readiness-webinar

第二场网络研讨会关注“保存计划的成功案例”，主要讨论“规划”会怎样促进一个数字保存项目的实施。来自北卡罗来纳大学、俄勒冈州立大学和奥比斯梯级联盟（Orbis Cascade Alliance）的发言人分享了他们的数字保存策略和做法。

在线观看请访问：

http://www.slideshare.net/DuraSpace/22112-preservation-planning-success-stories-webinar

第三场网络研讨会，关注“联盟数字仓储的保存和归档要点”，发言人是联盟数字保存库（Alliance Digital Repository，ADR）的主任Robin Dean。Robin通过一个案例学习让大家知道ADR合作体（包括科罗拉多联盟研究图书馆）是怎样将其数字典藏从Fedora系统转移到Islandora系统上的。Robin还参与讨论了ADR保存规划是怎样指导他们在迁移过程中的决策行为的。

编译自：

http://duraspace.org/recordings-available-%E2%80%9Cknowledge-futures-digital-preservation-planning-series

（么媛媛编译，李红培 吴振新校对）