国外数字资源建设热点及其给NSTL的启迪

吴波尔[1]，张建勇[2]，揭玉斌[3]，梁冰[4]
（1. 国家科技图书文献中心，北京 100038；2中国科学院文献情报中心，北京 100190；
3. 中国化工信息中心，北京 100029；4中国科学技术信息研究所，北京 100038）

摘要：通过访问Elsevier数据中心、汤姆森科技总部、ITHAKA的Portico系统和多伦多大学图书馆，介绍上述机构的发展战略、数据管理中心建设、数字资源长期保存系统建设和开放获取策略等事项，并相应地提出有关NSTL发展战略的几点思考。

关键词：发展规划；数据管理；长期保存；开放获取
中图分类号：G203
DOI：10.3772/j.issn.1673-2286.2015.04.001

近年来，数据管理、开放获取和数字资源长期保存，是数字图书馆发展规划中必须重点考虑的课题，为此，2014年国家科技图书文献中心（以下简称NSTL）组织专家团队前往美国和加拿大访问了Elsevier（爱思唯尔）数据中心、汤姆森科技总部、ITHAKA的Portico系统和多伦多大学图书馆，围绕数字资源发展战略、数据管理、长期保存等进行了深入地访谈和交流，探讨了数字环境下科技文献信息工作的发展方向，以期为NSTL制订发展战略提供参考。

1 数据管理中心成为建设重点

（1） Elsevier技术服务数据中心

Elsevier技术服务的数据中心位于俄亥俄州（Ohio）的代顿市（Dayton），建有网络服务设施、软件和通信设备等以支持Elsevier的产品和服务，是美国同类机构最大的数据中心之一。核心业务包括在线服务、数据管理和备份，以及应用开发、测试和业务管理等[1]，见图1。

基础设施包括空间、电力、环境和物理安全，每个部分都有监控设备。监控区实时监控网络、服务器和电力供应状况等各种运行数据，同时也显示美国和世界的形势变化并评估可能的影响。网络监控包括实时的网络运行状况监控，是否有拥堵和中断等，有双重的网络供应商的网络通道。平台服务包括企业计算平台：服务器、主机、存储、数据库和操作系统等。

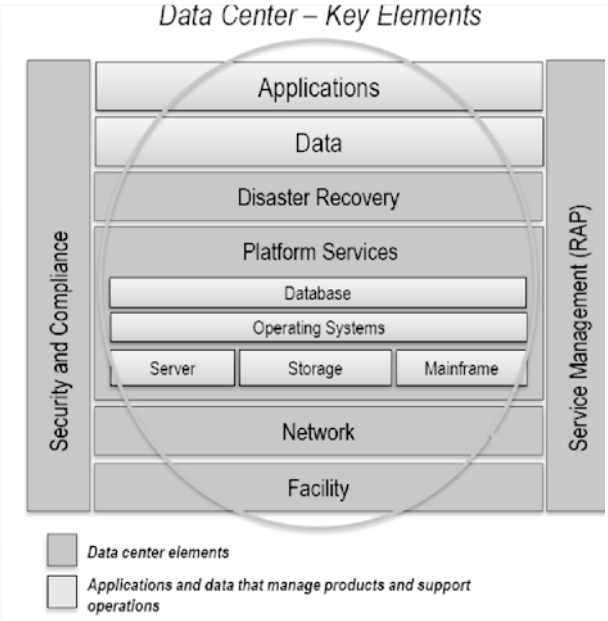

图1 Elsevier技术服务数据中心业务核心要素

灾难恢复是数据中心的核心业务，整个基础设施的设计能抵御天气和其他外在的风险。所有主要的系统都是完全的冗余设计，包括设备、电力、通讯等。所有

重要的数据实时备份到后备系统以确保数据丢失零风险。所有在线数据都会备份到磁带等多种介质，并且存放在不同地点，同时成为数据持续保藏计划的一部分。

冗余设计是该数据中心设计的特点。代顿市是重要的工业基地，有充足、可靠的电力供应，还设计了独立的柴油发电机和储存油罐，充足的柴油发电设备可以为数据中心供电10天，另外还有后备电池用于断电后切换到柴油发电期间的供电。计算机硬件和软件冗余体现在安装有同样的产品的硬件和系统有多套处于运行状态。网络冗余体现在所有重要的产品系统都部署双通路网络，外部网络接入多个网络提供商，即使1个或2个网络提供商出现网络阻断也不会影响网络链接。

数据中心的规划设计能保证在1天的数据恢复期间面向用户的联机服务不停顿，如果数据中心面临一天以上的完全停电状况，重要的服务不会停止。保证所有的系统和服务都能从较长时间的灾难中恢复。

（2）多伦多大学图书馆地图与数据图书馆

多伦多大学图书馆地图与数据图书馆（Map & Data Library，MDL）的馆藏资源包括：数字数据（Digital Data）、地理空间数据（Geospatial Data）和纸版地图。涉及地质、气候、土壤、水、人口、土地使用、经济、选举、交通、城市事务和历史等方面。还有关于图像的专业基础数据库、媒体共享视听图书馆、缩微资料馆藏，以及音频在线收藏。MDL的服务包括为本校师生、科研人员和社会公众查阅和使用地图、数字数据、地理空间数据和有关软件。MDL把数据服务（Data Services）划分为四个方面[2]：

- 数据采集（Data Acquisition & Collections），包括通过所参加数据联盟、许可协议和开放途径，采集各种微观数据（Microdata）和观测数据（Surveys），目前已采集10万多个数据对象；
- 数据存取和可利用（Access & Usability），推进数据开放计划、制订ODESI（Ontario Data Documentation, Extraction Service and Infrastructure）规范和标准、开发有关软件和分析工具等实现数据的可利用和数据的可视化；
- 数据统计与读写（Data & Statistical Literacy），包括对数据的存取、转换、操作、评估、汇总和描述，应用统计软件，精确的统计结果呈现和作为证据的数据分析、解读、评价等；
- 数据管理与长期保存（Data Management & Preservation），参加关于研究数据管理、保存及利用的各种发展计划和研究、建设项目，积极发挥图书馆在研究数据的管理、保存及利用中的社会责任和专业作用，并在研究数据保存、管理的拓展和实践中，努力开展与产生、拥有、掌握研究数据的科研人员的协作。

2 数字资源长期保存系统得以形成

数字资源长期保存是为确保内容在很长期限内仍具有持久的可用性、真实性、可发现性和可访问性而采取的一系列必要的管理政策和活动。数字保存的关键目标包括：可用性，即数据的知识内容必须始终可以通过当前技术的传送机制使用；真实性，即内容的出处必须可靠并且内容是原作的真实复制；可发现性，即内容必须具有逻辑书目元数据，以便最终用户以后仍能找到它；可访问性，即内容必须准备好，可供相应人群使用。此次访问重点关注了数字资源长期保存系统的发展情况，重点考察了业界比较有影响力的两个系统：第三方保存机构Portico的保存策略和系统、CLOCKSS（Controlled LOCKSS，管理多重副本以保护电子资源）的保存理念和系统构成。

Portico认为数字资源和印本资源相比有自己的特点：数字资源现在的使用方式是获得许可而非拥有；数字内容的使用依赖于技术；技术变革的快节奏使得其存在固有的易受损特性；有众多电子格式。提出图书馆和出版商如何保证各种机构对学术资源的长期访问，保存的责任应由谁承担，图书馆、出版商，还是第三方。数字资源的规模和复杂性让单个机构保存资源感到吃力，需要图书馆、出版商和第三方之间展开协作以提高成本效益。数字资源长期保存需要大规模的基础架构（技术、组织以及专业技能）和经济模型来支持其可持续发展。

Portico的目标是帮助图书馆和出版商安全可靠地从依赖印刷物过渡到依赖电子内容，保持与出版商达成存档协议以收集和保存内容，直接从出版商接收内容，保持与图书馆达成协议以支持存档。Portico建立了面向电子资源的“保险政策”，对图书馆的价值体现在：当某些已存档的内容丢失、绝版或被弃时，为图书馆提供对这些内容的访问权限（无论图书馆以前或当前是否订阅了该内容），包括出版商停止运营或出版商终止出版。对出版商的价值体现在降低（甚至是免去）出版商的内部存档成本；满足出版商对第三方存档的需求；将源文件转化为存档格式并方便未来进行格式转换[3]。

Portico在2011年已开始了电子书的保存服务，面向参与和非参与图书馆提供，此服务涵盖Portico保存的来自于当前和未来电子书出版商参与者的所有内容，访问情景模式反映当前访问模型，即“触发事件”，图书馆和出版商共同承担保存成本。未来计划保存图书馆内部产生的数字内容，与国家图书馆合作，与科研团体开展协作，研究动态内容的保存以及数字科研的新形式。到2013年，Portico系统中参与的出版商数量有涉及2000多家社团和协会的300家出版商，已承诺保存21,876种电子期刊、51万种电子书和122种数字集合。参与的图书馆数量达到922家[4]。

Portico的技术要点是通过收录系统ConPrep处理内容。ConPrep在Portico内容模型中提供存档单元，是建立在Documentum基础上的工作流系统。确保新工具（每个出版商一套新工具）能够被开发出来并置入工作流中。ConPrep产生的存档单元与存入其中的单元类似，但会出现一个新XML文件，这是规范为 JATS 标准的出版商XML文件版本。此外还会生成一个保存元数据文件(PMD)。Portico内容模型是一个源于PREMIS、METS、DIDL（JPEG-21标准）和经验的六层分级模型[5]。

CLOCKSS是一个分布式保存系统，2006年由几家全球大型学术出版商和著名研究图书馆共同建立的数字资源长期保存系统，共同承担保护学术电子资源的责任。其管理和合作模式具有共同管理的特点，出版商和参与图书馆在决定管理程序、发展重点及何时触发资源方面享有同等的权利。CLOCKSS现时为全球各委员会成员共同管理的非盈利机构。CLOCKSS保存系统界定的数字资源保存为通过可靠的流程和程序确保数字资源在未来的可用性。当技术环境发生变化

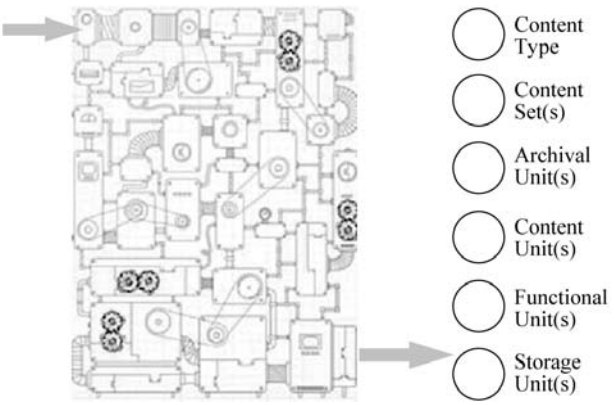

图2 六层分级模型

或停订等情形发生，可保证数字资源的可被检索利用。CLOCKSS的技术框架为分布式结构，利用LOCKSS技术建立了多个节点的保存网络，这些节点分布于全球12所大型研究图书馆，这些图书馆包括澳大利亚国立大学，德国柏林洪堡大学，美国印第安纳大学、OCLC、莱斯大学、斯坦福大学和弗吉尼亚大学，日本国立情报学研究所，意大利米兰圣心天主教大学，加拿大阿尔伯塔大学，英国爱丁堡大学和中国香港大学。

CLOCKSS系统重点关注数字资源，尤其是电子期刊和电子图书。根据有关统计，目前全球有超过5000家的STM类出版社，超过25000种经过同行评议的STM期刊，96%有电子版本。CLOCKSS系统主要为三类群体服务：科研人员、学生和最终用户（使用资源者），图书馆（购买和组织资源者），出版社（拥有资源者）。CLOCKSS确保资源使用者在资源不可获取时提供开放免费的服务，确保图书馆订购内容的长期有效可访问，确保出版社无法服务时责任的免除。电子图书的长期保存在技术上与电子期刊一致，但由于电子图书的版权归作者所有，所以需要管理更多的权利细节问题。虽然还要面临格式和平台的多样性等问题，但保存电子图书这个目标是确定的[6]。

多伦多大学图书馆是安大略省大学图书馆委员会OCUL的重要成员。OCUL集团联盟共有21个成员单位，通过Scholars Portal平台实现集团联盟成员文献资源的一站式检索，为集团成员单位提供各类文献资源的共享和联合参考咨询服务。OCUL集团联盟为电子资源管理和许可协议、用户权限的管理提供多种工具。OCUL集团联盟通过集团协议，以Local Hosting方式实现集团订购电子文献资源的本地长期保存，根据与供应商签署的Local Hosting协议，按照协议规定的授权许可的资源范围和集团许可用户范围通过本地保存和服务系统进行资源的管理和用户使用资源的权限管理，提供在线服务。同时，OCUL集团联盟通过Scholars Portal平台开展社会科学数据、地理空间数据等研究数据的管理和可信赖的长期保存服务[7]。

3 开放获取与深度加工成为新策略

Elsevier积极和全球的资助机构、大学和研究机构建立关系，支持金色（在开放获取期刊上发表）和绿色（作者将已发表的论文存入开放信息库）的开放获取，

帮助科研人员实现论文的获得。2010年Elsevier推出了第一份开放获取期刊，现已有39种开放获取期刊。目前，Elsevier和16家基金组织签署了金色方式的开放获取，可立即获得论文的获取权利，并且可自动地传递或存储到这些机构的仓储库。绿色方式可在具体期刊允许的滞后发布时间后提供检索获取，当前和7家机构签署了绿色开放获取的协议。有超过1500种期刊为复合开放获取期刊，提供开放获取论文的出版选择，大约40%的作者选择开放获取方式发表论文而费用由资助机构负责，只有很小部分的期刊出版不提供OA选择[8]。

Elsevier的Production &Hosting服务为本地的期刊提供审稿和发布平台搭载服务，帮助这些期刊成为高质量的国际期刊，目前有53种各国的期刊搭载到平台上，其中14种期刊来自中国。为消除信息鸿沟，Elsevier为公众以极低的价格提供250种刊的检索获取服务，并且以单篇0.99美元到3.99美元的价格提供服务。也为博士毕业尚未有研究岗位的人员提供ScienceDirector所有资源的全文服务。Open Archives提供80种刊的有条件获取服务。

汤森路透拥有用于研究评价和管理的全球标准。汤森路透科技在知识产权和科学方面的服务创新的生命周期为：从发现到发展再到商业化。发现：通过基础研究、合作和建设的循环，把全球最好的科学链接成一个科学社区，代表产品有Web of Knowledge、Endnote、Scholar One；发展：通过应用研究、发展和计划的循环，驱动更有效的和具有创新性的研究和发展，代表产品有Cortellis；商业化：商业化并且保护世界上最有价值的发明，代表产品有IP Solutions。在支持科学和学术研究方面，为客户提供内容、工具和服务用于激励发现、促进写作和指导关键的战略决策。Web of Science数据库是汤森路透的重要产品之一，它收录了12466种世界权威的、高影响力的学术期刊，内容涵盖自然科学、工程技术、生物医学、社会科学、艺术与人文等领域，最早回溯至1900年。其中，自然科学8595种，社会科学3105种和人文学科1723种。Web of Science选刊过程强调质量，每一种期刊都根据其所属学科领域的影响力筛选而出。此外，该数据库还跟踪各个会议录论文、会议汇总或系列会议的影响力，发现新趋势，以帮助研究者开展成功的研究并获得科研基金[9]。

Web of science引文检索数据库，收录了论文中所引用的参考文献，并按照被引作者、出处和出版年代编成独特的引文索引。并且把引文中各种数据表现形式归一化，以便于用户高效快速检索。通过参考文献即文献间的引证关系来展开检索，通过作者所引用的参考文献发现论文间潜在的科学关系，以获取相关的科学研究信息。通过回溯以往的研究成果并跟踪其最新进展，了解谁在引用研究者的论文，评估著作的影响力，并追踪当前最受关注的核心热点论文。对各种期刊和会议录文献进行向前回溯和向后追踪，将跨越时代、跨越学科的研究联系起来，以发现具有影响力的信息。

德温特世界专利索引（DWPI）是全球收录最全面的深加工专利数据库，覆盖了来自全球90多个国家的40余万个机构和组织，收录了来自世界各地超过4700万专利文献和超过2100万专利族，覆盖超过47个全球范围的专利局，包括主要的和新型的创新中心。

4 对NSTL规划设计的启示

本次访问围绕着数据管理、数据的长期保存和机构发展战略问题展开，通过交流加深了对数据管理中心的直观印象，了解了数据管理流程和策略，数据恢复系统框架和体系。特别是对国际上比较著名的数字资源长期保存系统进行访谈，对长期保存系统的建设过程和运作方式有了全面的认识，对NSTL制订发展战略有借鉴意义。

（1）强化自身数据中心的建设

本次访问的机构均以数据管理和服务见长，各个机构都将数据有效管理、组织、存储和服务作为本机构核心的业务。其中，Elsevier的数据中心是一个公司的最重要的数据中心，设立了多重的安全框架和流程保证公司重要的数据资源在遇到各种突发情况都能安全存储和服务；CLOCKSS和Portico则是重要的第三方数字资源长期保存服务机构，为出版商和图书馆提供可靠保存服务，不论技术和环境如何变化，都能提供可靠的服务；而多伦多大学图书馆则建立了国家层面的联合长期保存体系，注重数据的收集保存和服务；汤森路透科技也是建立了多种类型的数据库，形成了可靠的数据管理体系和评价体系。在当前数字资源占主导的情况下，NSTL作为国家级信息服务机构，建立一个国家级的数据中心将有力支持我国的科研进一步的发展。本次访问获得的数据中心物理设计、灾难恢复系统、管理机制和数据管理流程对建立NSTL数据中心具有重要的参考价值。

（2）促进数字资源长期保存

数字资源现在已经成为核心的资源，数字版本已经成为权威的版本，文献资源的E-only化已经在出版和图书馆领域成为主流。参加Portico和CLOCKSS的出版社已经超过200家，并在不断增加，参加两个保存体系的图书馆也分别达到了近千家。多伦多大学图书馆则联合加拿大其他高校建立了国家层面的长期保存和服务中心。数字资源的特点决定了长期保存的价值和重大战略意义。NSTL以印本为基础的业务体系面临挑战。全球出版资源不断加深数字出版和开放出版，国内众多的高校图书馆和研究型图书馆在逐步走向E-only化，资源和服务的网络化是当前的主流模式。NSTL作为国家级信息机构，应认识到这种紧迫的形势并面对这一挑战。参考已成功建设的数字资源长期保存体系的经验，建立国家层面的数字资源长期保存中心的任务历史性地落在了NSTL头上。而如何建设，Portico和CLOCKSS的第三方长期保存机构和多伦多大学图书馆的联合长期保存体系都给NSTL建设长期保存中心提供了重要的参考。

③ 加深国际间的协同和合作

国际间的科学研究协同和合作日益扩大已被文献计量证明，学科间的融合日益深化。文献信息领域的协同和合作也日益深化和扩展，数字资源的长期保存领域很好地体现了这一点，出版商和图书馆合作，第三方机构介入将出版商和图书馆紧紧地联系到了一起，各方共同努力实现共同的目标：保证资源在未来可用，保证资源在当前服务失效时仍然可用。在开放获取资源领域，资助机构、出版商、图书馆、用户紧密地联系到了一起，资助机构不断支持金色开放资源，出版商同意图书馆的绿色之路，用户既是服务者也是资源生产者，各方方面紧密协同，共同促进科学的传播和服务。NSTL在这个变化的环境中，应更多地融入全球高效协同的网络，成为其中重要的一份子，发挥重要的作用。

参考文献

[1] Lu A, Eric C. Elsevier Content Backup and Recovery Overview[R]. CHINESE DELEGATION VISITING ELSEVIER IN DAYTON OHIO,2013.

[2] Research Data Management[EB/OL]. [2015-03-23]. http://data.library.utoronto.ca/content/research-data-management.

[3] Wittenberg K. Portico：历史沿革、组织结构和业务模式[R]. CHINESE DELEGATION VISITING PORTICO IN NEW YORK,2013.

[4] Portico. Who Participates in Portico? [EB/OL]. [2015-03-23].http://www.portico.org/digital-preservation/who-participates-in-portico.

[5] Kirchhoff A. Metadata for Preservation: A Digital Object's Best Friend: Part 2: Implementation 2014 [EB/OL]. [2015-03-23]. http://www.portico.org/digital-preservation/wp-content/uploads/2013/12/NISO-PREMIS-20130213.pdf.

[6] Kiefer L. The CLOCKSS Archive: Challenges in Digital Preservation[R]. CHINESE DELEGATION VISITING ELSEVIER IN DAYTON OHIO,2013.

[7] OCUL. 2015 [EB/OL]. [2015-03-23]. http://ocul.on.ca/about.

[8] Alicia W. Funders, Libraries & Publishers:Working Together to Support Worldclass Research[R]. CHINESE DELEGATION VISITING ELSEVIER IN DAYTON OHIO,2013.

[9] MacGregor K, Testa J, Marie E M. Connecting the scientific community to the world's best research[R]. CHINESE DELEGATION VISITING SCIENTIFIC& SCHOLARLY RESEARCH IN PHILADELPHIAN2013.

作者简介

吴波尔，总会计师，国家科技图书文献中心（NSTL）副主任。
张建勇，研究馆员，NSTL数据研究管理中心副主任，通讯作者，E-mail：zhangjy@mail.las.ac.cn。
揭玉斌，研究馆员，中国化工信息中心副主任。
梁冰，研究员，NSTL网络管理中心副主任。

The Hotspot in International Digital Resource Development and its Effect to NSTL

WU BoEr, ZHANG JianYong, JIE YuBin, LIANG Bing
(1. National Science and Technology Library, Beijing 10038, China; 2. Library, Academy of Sciences, Beijing 100190, China;
3. China National Chemical Information Centre, Beijing 100129, China; 4. Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Paper introduces the development strategy of these organizations, data management center construction, the systems of long-term preservation of digital resources;open access policy and other such matters by visited the data center of Elsevier, ITHAKA ,Thomson Scientific and library of Toronto University. Accordingly puts forward some thoughts on NSTL development strategy.

Keywords: Development Strategy; Data Management; Open Access; Long-term Preservation

（收稿日期：2015-04-06；编辑：王立学）