《专题：数字资源长期保存》

韩国国家图书馆的网页存档系统

韩国国家图书馆 著　余敏 编译

1．引言

伴随着信息和通信环境的迅速发展，越来越多的人类知识成果可以在网络上以数字的形式被获取，然而这些数字资源却转瞬即逝。作为加工、管理和保存这些数字资源的长期方法，数字存档计划被认为具有永恒的价值。自从上个世纪 90 年代开始，不少国家，像澳大利亚、美国以及欧洲国家就在保存这些数字资源方面展开了持久的努力。而作为这些国家的长期规划，这些项目一般都是由相应的国家图书馆和其机构和组织共同合作开展的。

随着图书馆在数字信息环境下的地位变化，韩国国家图书馆（NLK）已经制定出一整套有效的国家信息服务机制，该服务旨在向全国人民提供有质量保证的数字化信息收藏和公共服务，并为下一代国民保存这些智力成果。

为了搜集内容多样的网页内容，以迎接 2008 年韩国国家数字图书馆的开放，NLK 已经开展了一项称之为网络存档与搜索因特网资源（OASIS，Online Archiving & Searching Internet Sources http://www.oasis.go.kr/）的计划，用于网络数字资源的搜索和保存。OASIS 系统于 2005 年 12 月开发，其目标在于为下一代国民保存数字资源，搜集并保存国家数字文化遗产，并建立起对数字资源的规范化管理政策。

作为韩国文化及旅游部（Korea's Ministry of Culture and Tourism）的四大创新品牌政策之一，OASIS 得到了政府的极大关注。在该项目的 05 到 06 年阶段，韩国政府投入了大约 100 万美元以支持对网络数字资源的搜集和保存。根据目前的中长期规划，从 2007 年开始，政府将会对那些系统化的发展方向提供更大的支持。同时，政府将针对性地提高预算，以支持 2008 年国家数字图书馆的开放和到 2010 年达到百万网页资源搜集的目标。

2．OASIS 网页资源搜索方法

2.1 有选择性的网页资源搜集

NLK 进行网页存档的基本方法就是选择性的搜集。目前主要搜集两种类型的对象：网站和其他单独网页数字资源。资源的有选择搜集遵循相应的已建立的馆藏发展政策。同时我们将逐渐把搜集的对象扩大到视频、图像和音频。

在那些具备潜在收藏价值的对象中，可能对应的印刷版已经存在，但目前我们将会根据馆藏发展政策继续搜集这些资源，而忽略其潜在的电子化可能。

2.2 OASIS 馆藏目标和馆藏政策

对目标资源的搜集主要基于以下几个因素：有效满足目前或者未来的信息需求，作者受欢迎程度，信息的稀缺性，学术内容，新颖性，更新的频率，以及信息可获得性。

要成为国家数字资源的保存对象，这些数字资源必须是被韩国民众认可的，与韩国的社会、政治、文化、宗教、科学或经济相关的、著作。并且，这些著作的著者必须是其专业领域的权威，例如在韩国大学里的知名教授或研究者。同时，他们必须在国家或国家层面有被认可的学科建树。

这方面的例子包括那些被认为在新颖性，稀缺性以及实用性上有搜集价值的数字资源的保存和搜集。例如对于目前热门的国会选举和新的建都地址。数字资源的搜集对象还包括那些被国际权威机构评估认可的期刊文章。

2.3 OASIS 搜集步骤

NLK 对网络数字资源的收集有如下 5 个步骤：

首先是对馆藏对象的评价过程。一个方法是通过搜集政策，另外一个则是通过一个由来自不同领域的专家所组成的数字资源搜集和保存委员会来进行。第二个步骤主要是处理那些选定搜藏对象的知识产权问题，并由 OASIS 系统进行相应搜集。第三步则是根据都柏林核心集的基本元素，如标题、URL、出版商或文摘以及主题分析等对搜集的信息资源进行编目。第四步是对编目工作的核查，以修正错误，并由学科专家根据资源的价值做出是否保存的决定。第五步则是资源的保存过程，这个过程中，已搜集的数字资源被转化成相应的保存格式，选取保存媒介，然后保存到相应的媒介载体中去。

第六步也是最后一个步骤则是，向用户提供那些已解决版权问题的网络数字资源馆藏。

2.4 OASIS 馆藏资源年度统计

馆藏的搜集开始于 2004 年，目前 OASIS 已拥有 156，798 个馆藏资源。馆藏容量大约为 2.4TB。

表格 1. OASIS 馆藏数据 (标题数)

资源类型	2004	2005	2006	合计
单独数字资源	43,861	45,280	42,958	132,099
网站	1,218	2,716	20,765	24,699
合计	45,079	47,996	63,723	156,798

单独的数字化资源是指那些由政府组织、其他公共机构、研究机构、协会以及个人所产生的文档文件。对于网站资源，我们搜集了包括新都地址，选举网站以及地方节日在内的各个主题领域的内容。我们的目标是在 2010 年达到 100 万网页资源存档，同时存档的对象将被扩展到视频、图像以及声音。

对于被搜集对象的知识产权许可申请方面，2005 年，在被要求给与许可的 1002 家机构当中，只有 209 个机构同意我们对其资源进行搜集和保存，授权率仅有 20%，而整个数据到了 2006 年只有 17%，即在要求的 650 家机构中只有 112 个机构同意授权。

由于缺乏对数字存档的共识，以及版权持有者对于知识产权的低授权率，应该积极地鼓励政府和其他主要组织以促进他们志愿参与到诸如数字存档计划中来。

3. OASIS 流程和过程

OASIS 的流程和过程设计是分别面向网站以及单独的数字资源的。

由于网站资源的内容处于不断的更新变化当中，因此网站资源的搜集过程并不是通过一个循环就能够解决的。这就有必要固定某个时间段来对其进行搜集和保存。然而，让一个管理员持续不断地对海量的网站资源进行监控是不可能的，那种认为应该无条件地在一段时间内，如每个月，每隔两个月或者六个月，对每个资源进行搜集以保存的做法是对资源的浪费。

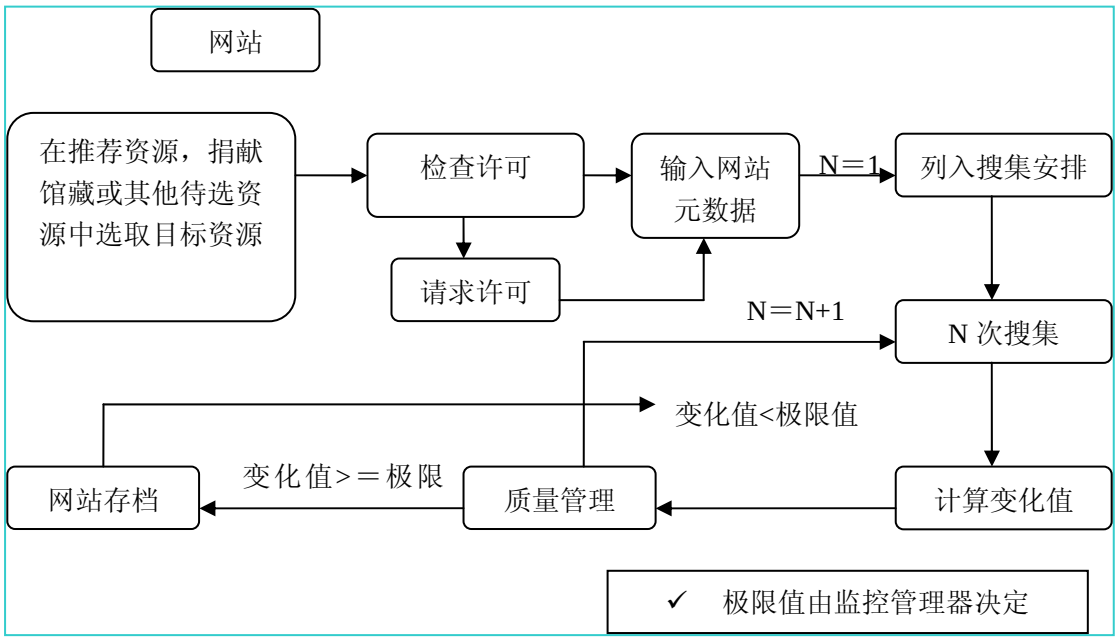

图 1.网站存档工作流程

OASIS 系统中使用专门的资源搜集机器人对已注册的网站资源进行搜集，并动态监控其变化，通过将目前的状态与先前保存的数据进行比较，给出一个变化的相应数值。管理器将根据这个数值来决定是否应该保存最新的网站数据。

总体的工作流程和过程如图标 1 所示。首先根据用户的推荐、作者的提供或者是管理器的自动选择，定义目标网站的基本信息以及资源搜集计划。网页机器人随后根据一个特定的搜集计划对网站的信息进行第一次镜像搜集。

管理器进行第一次搜集的检查，并制作相应的保存副本。之后，网页机器人按照计划进行第二次资源的搜集并给出与第一次搜集的资源的相应变化率。管理器检查变化率，决定是否应该制作二次收藏的备份。第三次则是与第二次搜藏的备份比较以获知相应变化率。

被选中的单独的数字资源是由机器人进行搜集的。机器人对目标资源进行搜集，检查副本，根据分类系统进行自动分类并进行摘要信息抽取。管理器为已经过前端处理的数字资源输入不同的元数据，检查并纠正从而制作出最终用于保存的编目数据。

4. 未来发展方向

由于知识信息资源呈现由印本向数字形式转变的趋势，在国家层面上对数字知识信息资源的搜集和保存的必要性开始得以重视。意识到数字资源存在的瞬间性，在 NLK 的领导下，作为一个国家性的工程，OASIS 系统正在对有价值的数字资源进行搜集以确保数字文化遗产传承到下一代。

为了完成这一使命，OASIS 提供了在未来数字环境下向机构提交网络电子资源的标准模型，同时，它还提供了网络电子资源搜集和保存的标准。

在资源的搜集、保存、管理以及公众服务方面的一些主要发展技术已经被应用到 OASIS 系统中。这些技术包括网页机器人代理的发展以及相应利用技术、自动分类技术、自动抽取技术、以及其他的搜集处理技术。在资源保存过程方面，需要完成对记录媒体的定期管理以及备份技术的相应研发。而在公众服务方面，应继续完善对于没有版权问题的资源的搜索技术。

作为即将于 2008 年开馆的国家数字图书馆的一个主要子系统，OASIS 系统将与 NLK

领导的相关组织进行合作以构建一个标准的分布式系统。该分布式系统将把网页资源的搜集过程深入到各个不同的学科领域。

编译自：A Web Archiving System of the National Library of Korea：OASIS .http://www.ndl.go.jp/en/publication/cdnlao/058/583.html. [2007-4-7]