웹 아카이빙 도구에 관한 연구

A Study on Web Archiving Tools

이성숙, 충남대학교 문헌정보학과 강사, inflee@hanmail.net

Lee Sung-Sook, Chungnam National University

<초록>

이 연구에서는 웹 아카이빙의 활성화를 위한 기초자료를 제공하기 위하여, 웹 아카이빙 관련 프로젝트에서 사용한 도구들을 살펴보았고, 웹 아카이빙 전용 SW 중에서 하비스팅 도구인 NEDLIB Harvester와 Heritrix, 접근도구인 Wayback Machine과 NWA Toolset을 중심으로 특징과 주요 기능을 검토하였다.

1. 서 론

웹은 학술커뮤니케이션뿐만 아니라 개인커뮤니케이션, 출판, 전자상거래, 마케팅 등 다양한 분야에서 중요한 위치를 차지하고 있으며, 정보원으로서 웹에 대한 의존도가 높아지고 있다.

그러나 웹이 갖는 가변성으로 인해, 웹페이지의 일부 혹은 전체가 변경되거나 사라지기 쉽다. 즉 아무리 중요한 웹페이지라고 하더라도, 그 중요성에 상관없이 새로운 페이지로 갱신되면 대다수 정보가 소멸된다. 이러한 상황에서 웹을 보존하는 웹 아카이빙의 중요성이 강조되고 있다.

웹 아카이빙은 시간이 지나도 진본을 유지하고 웹페이지에 접근할 수 있도록 안전하게 보존하는 활동이다. 웹 아카이빙의 필요성은 전통적으로 인쇄정보원의 수집과 이용, 보존에 중추적인 역할을 해왔던 국가도서관들이 먼저 인식했으며, 최초의 사례로는 1994년 캐나다 국가도서관의 EPPP(Electronic Publications Pilot Project)를 들 수 있다. 또한 비영리단체인 Internet Archive가 조직되어, 1996년부터 웹문서를 아카이빙하고 일반 이용자에게 서비스를 제공하고 있다.

이러한 웹 아카이빙의 사례가 증가하면서, 웹 아카이빙을 지원할 수 있는 전용도구의 개발이 필요하게 되었고, 몇몇 웹 아카이빙 프로젝트에서는 그러한 도구들을 개발하였다.

이 연구에서는 웹 아카이빙의 활성화를 위한 기초자료를 제공하기 위하여, 웹 아카이빙 관련 프로젝트에서 사용한 도구(Tools, SW)들을 살펴보았다.

웹 아카이빙 도구는 웹 아카이빙 접근방법과 밀접하게 관련 있기 때문에, 2장에서는 웹 아카이빙 프로젝트를 주체와 접근방법에 의해 범주화하고 사용된 SW를 살펴보았다. 3장에서는 웹 아카이빙 전용 SW 중에서 하비스팅접근을 위한 하비스팅 도구인 NEDLIB Harvester, Heritrix와, 접근도구인 Wayback Machine과 NWA Toolset을 중심으로 특징과 주요 기능을 검토하였다.

2. 웹 아카이빙 프로젝트와 도구

2.1 웹 아카이빙 관련 프로젝트

지금까지 여러 웹 아카이빙 프로젝트가 진행되었는데, 이를 주도한 기관에 따라 다음의 네가지로 범주화할 수 있다(Day 2003).

첫째 국가도서관의 경우이다. 대표적으로 스웨덴(Kulturarw3)과 호주(PANDORA)의 사례를 들

수 있고, 이외에도 미국, 영국, 일본, 프랑스, 독일, 덴마크, 노르웨이, 핀란드 등에서도 다양한 프로젝트가 추진되었다. 2003년 초에 실시된 조사를 보면 25개의 유럽연합의 국가도서관중 15개 도서관이 웹 아카이빙 관련 프로젝트를 추진하고 있는 것으로 나타났다.

둘째, 국가기록보존소의 경우를 들 수 있는데, 이는 웹이 증거문서로서의 가치가 있다는 인식에서 비롯되었다. 예를 들면 미국국립기록관리청과 영국국립기록보존소에서는 정부 웹사이트를 스냅샷으로 아카이빙하고 있다.

셋째, Internet Archive를 들 수 있는데, 이것은 1996년부터 운영되고 있으며 수집된 장서는 Wayback Machine을 통해 일반 이용자에게 제공된다.

넷째, 대학과 학계의 사례를 들 수 있다. 예를 들면 코넬대학교의 정치커뮤니케이션 웹 아카이빙 프로젝트를 들 수 있다.

2.2 웹 아카이빙의 접근방법

여러 기관에서 진행한 웹 아카이빙 프로젝트의 접근방법을 크게 나누면, 제출접근, 선택접근, 자동 하비스팅 접근의 세가지로 나눌 수 있다(Day 2003).

첫째, 전통적인 제출접근은 저장소에 웹사이트의 사본 혹은 스냅샷을 제출하도록 하는 것이다. 미국국립기록관리청에서 연방정부 웹사이트를 아카이빙할 때 사용하였다.

둘째, 선택접근은 보존할 개별 웹사이트를 선정기준에 의해 선정하고, 웹사이트 소유자의 사용허락을 받은 후, 미러링(mirroring) SW를 이용해서 수집하는 것이다.

셋째, 자동 하비스팅접근은 선택접근과는 달리 하비스팅 도구를 이용하는 것으로, 대개 특정 국가(혹은 전세계)의 모든 웹페이지를 크롤러가 링크를 따라가며 다운로드하여 수집하는 것이다. 이러한 접근을 사용하는 이유는 미래의 관점에서 자료의 가치를 판단하는 것은 어려운 문제이며, 선택접근시 선정작업에 너무 많은 인력이 필요하기 때문이다. 또한 비용측면에서 컴퓨터 기억장치의 가격이 급락하기 때문에 자동 하비스팅접근 방법을 사용하는 것이 가능하게 되었다. 자동 하비스팅접근에서는 특정국가의 웹페이지를 모두 수집하기 위해 국가도메인, 서버위치, 언어 등을 고려한다.

위의 접근방법들은 상호보완적인 관계라고 할 수 있다. 예를 들면 프랑스국가도서관은 하비스팅접근과 선택접근을 함께 사용하였다. 웹사이트의 갱신 여부와 중요성을 조사하기 위해서는 하비스팅 접근을 사용하였고, 딥웹(deep web)의 처리를 위해서는 선택 접근을 사용하였다.

2.3 접근방법에 따른 웹 아카이빙 도구

웹 아카이빙의 선택접근과 하비스팅접근을 위해서 다양한 SW들이 사용되었다. 우선 선택접근의 사례에 해당하는 호주국가도서관의 PANDORA, 미국국회도서관의 Minerva, 영국 국가도서관의 Britain on the Web, 영국 웹 아카이빙 컨소시엄(UKWAC) 등에서는 HTTrack이라는 미러링 SW를 사용하였다.

웹 아카이빙에 하비스터(harvester) 기술을 이용하려는 아이디어는 1996년에 스웨덴에서 생겨났다. 스웨덴 국가도서관은 Kulturarw3 프로젝트에서 Combined Harvester의 수정버전을 사용하였고(Lunds Universitets Bibliotek NetLab 1998), 이 SW는 AOLA 사례에서도 사용되었다. 또한 Internet Archive에서는 Alexa Crawler를, 프랑스국가도서관에서는 Xyleme Crawler를 사용하였다(<표 1> 참고).

웹 아카이빙 관련 프로젝트가 증가하면서, 웹 아카이빙을 위한 전용 하비스터의 개발 필요성이 대두되었다. 일반적인 색인로봇은 주로 웹검색을 위해 설계되었기 때문에, ①아카이브의 전체모듈 중 몇몇 모듈은 지원하지 않으며, ②인터넷 이미지를 가져오는 우선순위 등 몇몇 기능들은 적절하지 않고, ③문제있는 URL에

구분	웹 아카이빙 도구(개발기관)	사용 기관(프로젝트명)
하비스팅 도구	Combine Harvester(NetLab at Lund University Library)	스웨덴국가도서관(Kulturarw3) 오스트리아국가도서관(Austrian On-Line Archive)
	Alexa Crawler(Alexa)	Internet Archive
	Xyleme Crawler(Xyleme)	프랑스국가도서관(BnF Web Archiving initiative)
	NEDLIB Harvester(Finnish IT Center for Science: CSC)	노르웨이국가도서관(PARADIGMA) 체코공화국국가도서관(WebArchiv) 핀란드국가도서관(EVA) 리투아니아국가도서관(Archive of Electronic Resources)
	Heritrit(Internet Archive)	Internet Archive Nordic 지역의 국가도서관
접근 도구	NWA Toolset(Nordic Web Archive)	Nordic Web Archive에 참여한 덴마크, 핀란드, 아이슬란드, 노르웨이, 스웨덴 국가도서관
	Wayback Machine(Internet Archive와 Alexa Internet)	Internet Archive

<표 1> 자동 하비스팅접근에 사용된 하비스팅도구와 접근도구

대한 의도적인 건너 띄기 등 완벽하지 않은 수집방법은 웹 아카이빙에는 적합하지 않기 때문이다(Hakala 2001). 웹 아카이빙의 기술은 새로운 것은 아니지만, 완벽한 기능이 요구된다.

이러한 필요에 의해 NEDLIB(Networked European Deposit Library)에서는 웹 아카이빙 전용 하비스터로 NEDLIB Harvester를 개발하였고, NEDLIB Harvester는 유럽을 중심으로 한 다수의 국가도서관에서 사용하였다.

또한 Alexa Crawler를 사용하던 Internet Archive는 대규모이면서 완전하고 쉽게 커스터마이즈할 수 있는 웹 아카이빙 SW인 Heritrix를 개발하였고, 이 SW는 Nordic 지역의 국가도서관에서 사용되었으며, 향후 그 사용이 증가할 것으로 예상된다.

한편 하비스팅 SW에 의해 수집되고 저장된 웹페이지를 색인하고 접근할 수 있는 접근도구(Access Tools)도 개발되었다. Wayback Machine은 Internet Archive에서 사용되었고, NWA Toolset은 Nordic Web Archive에 속한 북유럽국가도서관에서 사용되었다.

다음 절에서는 <표 1>에 언급된 SW 중에서 특징적인 도구들을 선정하여 살펴보았다. 하비스팅도구로는 NEDLIB Harvester와 Heritrix를, 접근도구로는 Wayback Machine과 NWA(Nordic Web Archive) Toolset을 중심으로 살펴보고자 한다.

3. 웹 아카이빙 도구의 분석

3.1 하비스팅 도구

1) NEDLIB Harvester

NEDLIB Harvester는 자동 하비스팅접근을 위해 개발된 SW이다. 앞에서 언급하였듯이 선택접근에서는 주로 HTTrack이 사용되었는데, HTTrack은 특정 웹사이트의 URL을 입력하고, 웹페이지에 연결된 모든 구성요소를 다운로드 받을 수 있는 SW로서, 설치하기가 쉽고, 윈도우즈와 유닉스 시스템 모두에서 사용가능하다. 그러나 사이즈가 크거나 구성요소가 많은 웹사이트는 서버측에 심각한 대역폭 점유를 가져올 수 있다(http://www.httrack.com/). HTTrack은 선택접근의 소규모 사이즈의 수집에는 적합하지만, 수집하기 원하는 URL을 일일이 입력해야 되기 때문에 대규모 하비스팅에는 적절하지 않았다(Marill et al. 2004).

NEDLIB Harvester(http://www.csc.fi/sovellus/nedlib)는 대규모 자동 하비스팅 SW의 필요성이 제기된 이후 개발되기 시작하였다. NEDLIB 프로젝트의 일환인 NEDLIB Harvester의 개발은 1997년부터 2000년까지 네덜란드 국가도서관이 주도하였고, 개발기관은 CSC(Finnish IT center for science)이다. NEDLIB Harvester의

개발목적은 웹 아카이빙이라는 목적에 맞게 대규모로 웹문서를 수집하고, 시스템에 저장하기 위한 SW를 만드는 것이다.

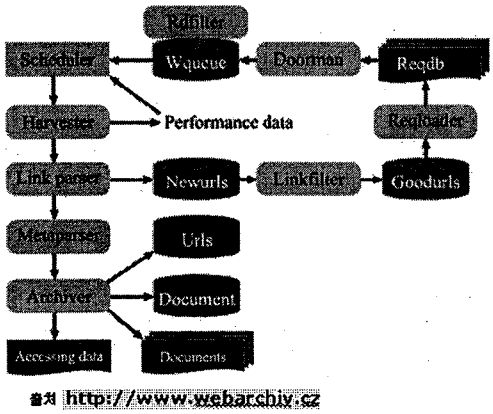
NEDLIB Harvester는 주로 납본활동의 일환으로 웹문서를 수집하는 국가도서관을 위한 것이지만, 다른 기관들에서도 사용가능하다. NEDLIB Harvester의 첫 번째 버전은 2000년 1월 공개되었고, 테스트를 거친 후 2000년 9월 버전 1.2.2가 발표되었다. 이 SW는 오픈 소스로서 공용도메인에서 사용가능하며, 다른 검색엔진과 이용자 인터페이스와 쉽게 통합되도록 설계되었다.

NEDLIB Harvester는 관계형 DB에 기반하며, HTTP/1.1과 HTTP/1.0, FTP-프로토콜을 지원하고, 일반적으로 로봇배제표준을 따르지만 납본 활동의 맥락으로 접근하기 때문에 이는 선택사항이다.

대개 웹 하비스터의 작동과정은 비교적 간단하다. ①먼저 접근할 URL을 주기 위해 일련의 웹문서 링크 세트를 먼저 부여한다(세트의 규모가 크면 클수록 좋다). ②해당 URL의 내용을 가져온다. ③내용을 분석하여 URL과 메타데이터를 추출하고, 저장한다. ④발견된 관심있는 URL을 선정하고, 스케줄에 추가한다. ⑤URL에 관한 정보를 주기하고, 위의 과정을 반복한다. NEDLIB의 전체적인 작동과정은 <그림 1>과 같다(Rissanen et al. 2001).

이러한 구성요소 중에서 특징적인 것은 'Scheduler'이며, 설계시 가장 큰 비중을 차지한 부분이다. 예를 들면 하나의 요청사이에 5초와 같이 고정으로 최소시간간격을 정의하는 것은 간단하지만 효과적이지 못한 경우도 있다. NEDLIB Harvester에서는 인라인 이미지와 같은 인라인 콘텐츠에 우선순위를 두면서, 일차 수집된 웹페이지의 통계를 근거로 스케줄링 알고리즘을 만들었다.

또한 웹 아카이빙을 위한 중요한 고려사항은 저장의 문제이다. 아카이빙 도구는 하비스팅 활동을



<그림 1> NEDLIB Harvester의 구성요소

통제하는 DB, 수집된 문서처리(예: 메타데이터 추출)와 아카이빙 준비를 위한 워크스테이션, 아카이브된 문서를 처리할 수 있는 충분한 디스크 공간이 필요하다. NEDLIB Harvester는 이진데이터의 블랍으로 된 HTML 파일을 한 번에 검색할 수 있다(Hakala 2004).

NEDLIB Harvester에서 사용된 아카이빙 관련 메타데이터로는 아카이브된 웹문서의 식별자(MD5 checksum), 문서위치, 시간 스탬프 등이다. 이러한 메타데이터가 아카이브된 웹문서와 함께 저장된다.

NEDLIB Harvester는 웹 아카이빙을 위해 개발되었기 때문에, 하비스팅 성능면에서 만족스러운 결과를 나타냈다(Marill 2004). 그러나 보다 복잡한 환경에서의 융통성이 부족하고, 사용자 인터페이스가 사용자 지향적이지는 않다. NEDLIB은 NEDLIB Harvester를 갱신하는 대신 IIPC(International Internet Preservation Consortium)에 참여하여, 다음에 소개될 Heritrix 개발에 다른 국가도서관과 함께 참여하였다(Rissanen et al. 2001).

2) Heritrix

Heritrix(http://crawler.archive.org/)가 개발되기 전 Internet Archive에서는 Alexa Internet에서 기부한 크롤러를 사용하여 웹페이지를 수

집·저장·색인하였다. 그러나 Alexa Crawler는 Alexa Internet이 소유한 SW와 기술을 사용하는데, Internet Archive 혹은 다른 기관에서 이러한 SW의 기술을 사용하거나 확장할 수는 없었다.

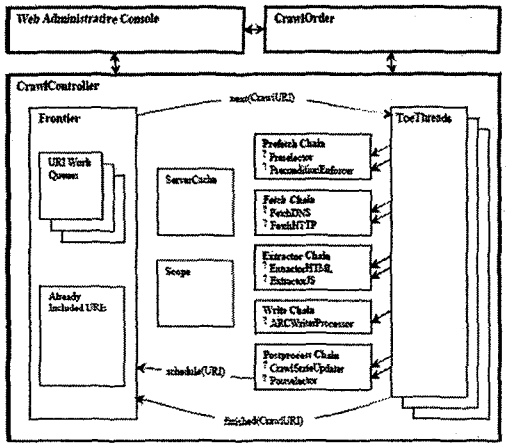
Internet Archive는 웹 아카이빙이라는 고유의 목적에 맞는 크롤링(crawling)을 원했고, 새로운 방법으로 웹을 크롤링하고 아카이브하기 위해 다른 기관(예: 각 국의 국가도서관)과의 협력이 필요했기 때문에, 대규모이면서 완전하고 쉽게 커스터마이즈할 수 있는 크롤러가 필요하다는 결론을 얻었다. 그러나 Internet Archive는 기존의 SW 중에서 광범위하게 크롤링할 수 있는 융통성 있는 SW를 발견할 수 없었기 때문에, 2003년 Heritrix의 개발을 시작하였고 2004년 8월 버전1.0.0을 발표하였다.

Heritrix는 오픈 소스이며, 약소 일반 공중 사용 허가서(Gnu Lesser General Public License, LGPL)하에 라이센스되어 있다. Heritrix는 Java를 소프트웨어 개발언어로 사용하였고, 로봇배제표준을 준수한다.

기존 크롤러와 비교해서 Heritrix는 대규모로 높은 대역폭을 갖추고 광범위하게 크롤링하며, 선정 사이트 혹은 토픽을 완벽하게 포괄하여 집중적으로 크롤링한다. 또한 각 URI를 단지 한번 다운로드하여 자원을 스냅샷하는 전통적인 크롤러와는 달리, 적절한 방문횟수를 관리할 뿐만 아니라 이미 가져온 이전 페이지 중 변화가 있는 페이지를 연속적으로 방문하는 크롤링이며, Internet Archive와 다른 기관들이 크롤링 기술 실험에 자유롭게 사용할 수 있는 실험용 크롤링이기도 하다(Mohr et al. 2004).

Heritrix의 작업절차는 위에서 설명한 일반적인 웹 크롤러의 과정을 따르며, Heritrix의 구성요소는 <그림 2>와 같고, 그 중 Scope와 Frontier, Processor Chains가 중요한 구성요소이다.

Scope는 주로 URI를 어떤 부분에 입력하고 출



<그림 2> Heritrix의 구성요소

력할지의 여부, 시작할 URI의 부여, 발견된 URI를 스케줄에 포함시킬지의 여부를 결정한다. Frontier는 URI의 수집 스케줄과 이미 수집된 URI를 추적한다. 또한 다음 방문할 URI의 선정과 이미 스케줄된 URI의 재스케줄을 방지하는 작업을 주로 한다.

Processor Chains는 Processors 모듈을 포함하는데, URI의 내용을 가져오며 결과를 분석하며 새롭게 발견된 URI를 Frontier에게 되돌려준다. Processors 모듈은 ①CrawlURI에서 지체, 재명령, 거부에 관한 정보를 받는 'Prefetch', CrawlURI에서 언급된 자원을 가져오는 'Fetch', 가져온 자원(예: HTML, CSS, JavaScript, PDF, MS WORD, FLASH의 형식 지원)에서 새로운 URI를 추출하는 'Extract', 크롤링 결과를 저장하는 'Write', 마지막으로 크롤-유지 작업(예: Scope에 포함되지 않은 URI의 테스트, 크롤러 의 내부정보 갱신 등)을 하는 'Postprocess' 등 5가지 체인으로 구성된다. Heritrix 크롤러는 많은 URI을 처리하기 위한 다중스레드(multithread)를 지원하며, 개별 작업 각각의 스레드를 ToeThread라고 한다(Mohr et al. 2004).

Heritrix 1.0.0의 주요 특징은 다음과 같다.

①하나의 크롤을 실행함으로 복수 웹사이트에

서 반복적으로 웹페이지를 쉬지 않고 수집한다. ②선정된 URI에 관한 수집 명령을 위해 먼저 넓이우선탐색을 수행할 수 있다. ③앞에서 언급된 Heritrix의 주요 구성요소는 확장성이 높다. ④다양한 옵션이 제공되어 환경설정이 용이하다. ⑤아카이브 파일, 임시 파일 등의 출력위치를 지정할 수 있다. ⑥다운로드 할 최대 바이트, 페이지, 크롤링 소요 시간 등을 지정할 수 있다. ⑦작업할 크롤링 스레드의 수를 지정할 수 있다. ⑧사용할 대역폭을 조절할 수 있다. ⑨작업 요청간의 최소/최대 시간의 구성이 용이하다. ⑩URI의 깊이, 링크 홉 수 필터(link hop count filters)를 포함한 배제/포함 필터링 기능을 제공한다. ⑩작업단계마다 다양한 리포트 기능을 제공한다.

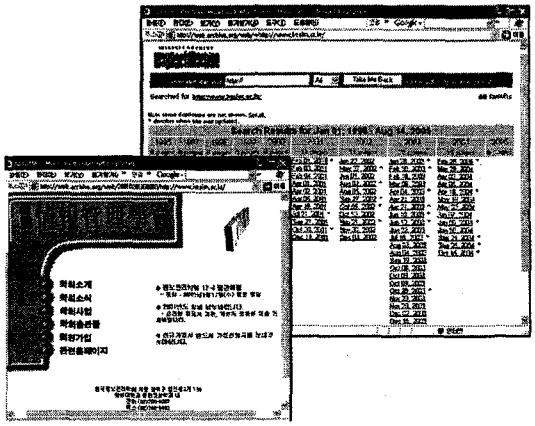
Heritrix는 가장 최근에 개발된 웹 아카이빙 크롤러로서, 그 동안 웹 아카이빙 크롤러에 필요한 많은 부분을 해결했다는데 그 의의가 있다. 그러나 대규모로 전환하기 위해서는 정교한 오퍼레이터가 필요한 문제, 리눅스만을 지원하는 문제, 더 많은 성능 평가의 필요 등은 해결해야 할 과제로 남겨두고 있다(Mohr et al. 2004).

3.2 접근 도구

1) Wayback Machine

Wayback Machine은 비영리 단체인 Internet Archives(http://www.archive.org)와 Alexa Internet (http://www.alexa.com)이 공동으로 개발한 시스템으로 아카이브된 웹페이지의 접근도구이다(Day 2003).

Internet Archive는 DNS(Domain Name System)를 통하여 URL을 획득하고, 획득한 URL에 대하여 날짜별로 페이지의 내용을 저장하는 방식으로 이루어져 있기 때문에, 이용자는 텍스트가 아닌 URL에 의해 검색해야 한다. 확장검색에서는 날짜, 파일형식(예: Image, Audio, Video, Binary, Text, PDF 등)의 검색



<그림 3> Wayback Machine의 정보관리학회 (http://www.kosim.or.kr/) 검색결과 화면

제한을 줄 수 있고, 아카이브된 웹사이트를 비교할 수도 있으며, 상세검색결과를 PDF 형식으로도 볼 수 있다.

이용자는 특정 URL을 입력하면, 약 1 페타바이트(petabyte)의 데이터를 검색하여 그 중에서 <그림 3>과 같은 간략결과화면을 얻을 수 있다. <그림 3>은 'http://www.kosim.or.kr/'이 아카이브된 시기별 리스트와 그 중에서 '2001년 2월 1일'에 아카이브된 웹사이트를 선택한 결과이다. 또한 특정 장서(예: Web Pioneers)에 대한 브라우징이 가능하다.

Internet Archive의 Wayback Machine은 일반 검색엔진과는 달리 특정 웹에 대한 시기별 웹사이트를 검색할 수 있다는 점에서, 역사가, 연구자, 학자, 사업가, 디자이너 등 다양한 이용자들에게 유용하다.

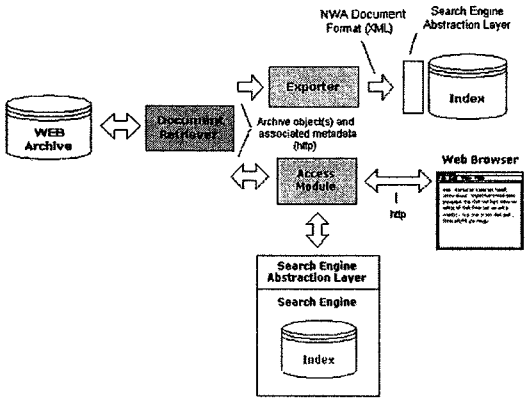
그러나 Wayback Machine에서는 전문검색에 의한 텍스트 검색이 어렵고, 모든 웹페이지가 완벽하게 아카이브하는 것은 아니기 때문에(예: Robots.txt, Javascript, Server side image maps 등), 검색결과 중 일부에 에러가 발생한다.

2) NWA Toolset

덴마크, 핀란드, 아이슬란드, 노르웨이, 스웨덴

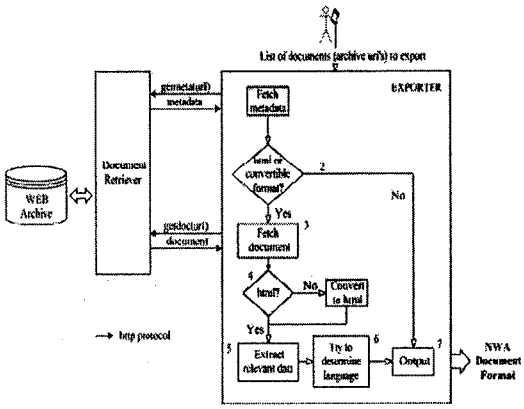
의 국가도서관들은 NWA(Nordic Web Archive)를 출범시켰다. NWA는 FAST와 Jakarta Lucene 검색엔진을 전문 텍스트 색인기로 사용하여, 아카이브된 웹문서를 위한 접근도구로서 NWA Toolset을 2000년부터 2002년까지 개발하였다. 이 NWA Toolset의 개발 목적은 아카이브된 웹의 내용을 일반적인 형식으로 변환하고, 텍스트 검색을 위한 색인어를 추출하고, 검색과 다른 시기의 URL간의 네비게이션이 가능한 인터페이스를 지원하는 것이다.

이 Toolset의 운영체제는 모든 POSIX (Linux/BSD/UNIX-like OSes)를 지원하며, 프로그램 언어는 Java, Perl, PHP이다. HTTP와 XML과 같은 공개표준을 이용하며, 표준 웹브라우저를 통해 접근할 수 있고, 오픈 소스이며, 많은 검색엔진과 저장 도구 및 형식과 통합될 수 있다. 특히, 이 SW는 원래 NEDLIB Harvester의 결과를 디스플레이 하기 위해 설계되었지만, Heritrix와의 호환도 용이하다.

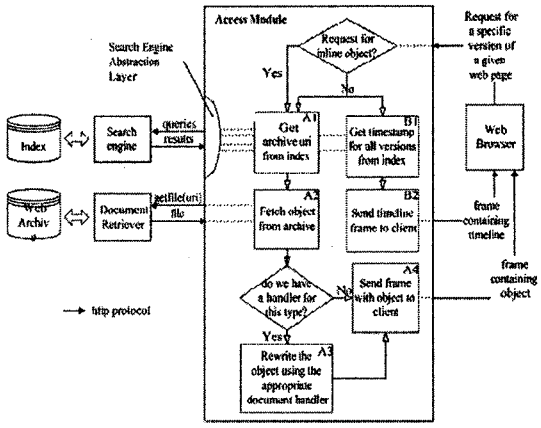


<그림 4> NWA Toolset의 구성요소

NWA Toolset은 <그림 4>와 같이 Document Retriever, Exporter, Access Module, Search Engine 등 4개의 주요 부분으로 구성된다(Hallgrímsson and Bang 2003). Exporter는 <그림 5>와 같고, Access Module은 <그림 6>과 같다. Access Module은 아카이브된 웹페이지를 검색하고, 브라우징하고, 항해할 수 있는 인터페이스를 제공한다(Bang et al. 2004).

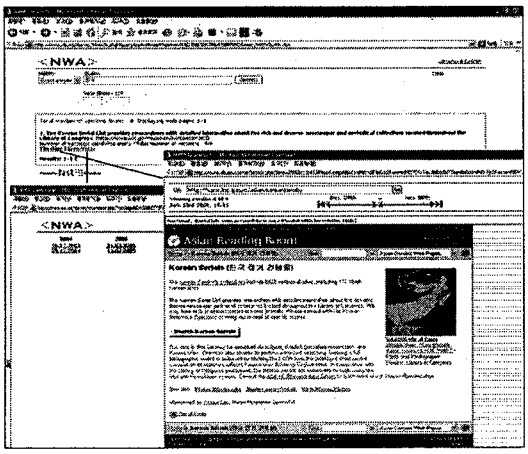


<그림 5> NWA Toolset의 Exporter



<그림 6> NWA Toolset의 Access Module

이용자가 특정 질의어를 입력하면, <그림 7>과 같은 결과를 얻는다. 이 화면은 '한국'이라는 질의어에 대한 검색결과이며, <그림 7>에서 'timeline'을 선택하면 웹사이트의 히스토리를 보여주며 아카이브된 시간에 의해 웹사이트의 히스토리를 따라갈 수 있다. 'overview'를 선택하면 아카이빙된 웹사이트의 리스트를 통해 그 히스토리를 알 수 있다. 확장검색에서는



<그림 7> NWA Toolset의 '한글'에 대한 검색 결과

불리언 연산자를 사용하는 본문검색, URL 검색을 포함하며, 표제, 주제, 링크와 같은 특정 필드제한과, 언어, 파일형식, 날짜 등도 제한하여 검색할 수 있다.

NWA Toolset을 이용하면, 아카이브된 웹페이지에 대해 현재 이용자들이 사용하고 있는 일반 검색엔진과 유사한 텍스트 검색이 가능하고, 동일 URL의 시기별 웹페이지의 항해가 가능하다.

4. 결 론

이 연구에서는 웹 아카이빙에서 사용된 SW에 대해서 살펴보았고, 하비스팅접근을 위한 도구로서 개발된 NEDLIB Harvester, Htrixtrix, Wayback Machine, NWA Toolset을 검토하였다.

국외에서 다수의 웹 아카이빙 프로젝트가 진행되고 있는 상황에서, 국내에서도 웹 아카이빙의 중요성을 인식하여 다양한 연구와 프로젝트들이 시작되었다. 예를 들면 국립중앙도서관의 웹로봇, 다음세대재단의 웹 아카이빙 시스템을 들 수 있다.

앞으로 웹 아카이빙을 위해서는 전용 SW로 개발된 Htrixtrix와 NWA Toolset의 활용방안을 모색하는 것이 바람직할 것이다. 이러한 도구들은 오픈 소스로서, 이 도구들을 개발한 기관들은 웹 아카이빙에서 '협력'의 중요성을 인식하고, 웹 아카이빙에 관심이 있는 기관의 상호협력을 꾀하고 있다.

향후 웹 아카이빙과 관련하여 이미 진행된 프로젝트에서 남겨두었던 문제들(예: 아카이빙된 문서의 전용 색인기, 딥웹(deep web), 새로운 웹-기반 기술, 인증이 필요한 웹)에 대한 연구가 필요하며, 지금까지 웹 아카이빙 관련 프로젝트들이 자원의 수집에 초점을 두었다면 이제 아카이빙된 웹페이지의 장기간 보존 문제에 관심을 기울어야 할 것이다.

웹 아카이빙을 위해서는 이러한 기술적인 문제 이외에도 웹 아카이빙 전략, 장서 정책, 법적인 문제(예: 납본법, 저작권법, 개인정보법 등)의 해결도 함께 이루어져야 한다.

참고문헌

Bang, S., K. Persson, and J. E. Halse. 2004. *NwaToolset Manual.* [cited 2005. 5. 22].<http://nwa.nb.no/demo/manual/manual.pdf>.

Day, M. 2003. *Collecting and preserving the World Wide Web.* [cited 2005. 5. 27]. <http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf>.

Hakala, J. 2004. "Archiving the Web: European experiences." *Program: electronic library and information systems,* 38(3): 176-183. [cited 2005. 4. 5].<http://images.emeraldinsight.com/emerald/pdfs/awards2005/aslib.pdf>.

Hakala, J. 2001. "Harvesting the Finnish Web space-practical experiences." *ECDL Workshop.* September 8, 2001. Darmstadt, Germany.

Hallgrímsson, Þ., and S. Bang. 2003. *Nordic Web Archive.* [cited 2005. 6. 20]. <http://nwatoolset.sourceforge.net/docs/nwa@ecdl2003.pdf>.

Lunds Universitets Bibliotek NetLab. 1998.

[cited 2005. 3. 13].<http://www.lub.lu.se/netlab>.

Marill, J. L., A. Boyko, M. Ashenfelder, and L. Graham. 2004. "Tools and Techniques for Harvesting the World Wide Web." *JCDL'04*, June 7-11, 2004, Tucson, Arizona, USA.

Mohr, G., M. Stack, I. Ranitovic, D. Avery, and M. Kimpton. 2004. *An Introduction to Heritrix.* [cited 2005. 5. 25]. <www.iwaw.net/04/proceedings.php?f=Mohr>.

Rissanen, M., J. Hakala, and K. Kaunonen. 2001. *Manual for installation and usage of the NEDLIB Harvester.* [cited 2005. 7. 30] <http://www.csc.fi/sovellus/nedlib/ver122/documentation122.doc>.