웹아카이빙 도구 비교분석 연구

Comparative Analysis of Web Archiving Tools

김희정, 국제백신연구소 정보자료실, heejung@ivi.int

Heejung Kim, Library and Information Service Center, International Vaccine Institute

디지털 자원의 장기보존을 위한 기법과 전략은 지속적인 관심 속에서 개발되어 오고 있다. 특히, 웹 자원에 대한 의존도가 증폭될수록 웹 아카이빙에 대한 중요성이 커지고 있다. 본 연구에서는 IIPC에서 제시하는 웹 아카이빙 체인의 네 단계에 해당하는 각 단계별 웹 아카이빙 툴과 그 특성을 살펴보았다. 대상이 되는 웹 아카이빙 도구는 총 9개로서, Heritrix, DeepArc, Web Curator Tool, NetarchiveSuite, BnFArcTools, Wayback, NutchWAX, WERA 그리고 Xinq 등이다.

1. 서론

웹 아카이빙은 가치 있는 웹 정보자원을 수집하여 장기 보존함으로써 후속 세대의 과학자, 역사가 그리고 일반 이용자들에게 지식을 전승하고자 하는 데에 그 목적이 있다.

웹 아카이빙에는 깊이는 얕지만 가능한 한 방대한 범위의 웹사이트를 수집하는 방식의 포괄적 아카이빙(extensive archiving)과, 아카이빙의 대상이 되는 웹 사이트 범주는 좁지만, 그 깊이를 최대한 수집하는 선택적 아카이빙(intensive archiving)의 두 방식이 있다.

특히, 아카이빙 대상이 되는 웹 자원의 방대함으로 인하여, 웹 크롤러(web crawlers)를 활용한 수집전략이 지속적으로 개발되고 있다.

본 연구에서는 웹 아카이빙을 위한 도구 중 해외 국가도서관 및 국가기록관 등에서 활발히 사용되고 있는 도구들을 대상으로, 웹 아카이빙 기본 단계를 적용하여 주요 내용을 정리하였다.

2. Web Archiving 과정

Masanés(2006)에 의하면 웹 아카이빙 과정은 다섯 개의 단계를 거치게 된다. 이에 대한 과정은 다음 <그림 1>과 같이 나타낼 수 있다(Kim and Lee 2007).

또한, 웹 아카이빙과 관련하여 국제적인 대표 기관으로 거론될 수 있는 IIPC (International Internet Preservation Consortium)에서는 웹 아카이빙 체인(web archiving chain)이라는 이름으로 기본적인 웹 아카이빙 과정 및 내용을 네 단계로 기술하고 있다.

Masanés(2006)에서 제시한 다섯 단계의 과정이 웹 아카이빙 관리자 차원에서 파악해야 할 전체적인 Cycle을 나타내는 것이라면, IIPC에서 제시하는 네 단계는 아카이빙 프로세스 자체에 초점을 맞춘 것으로서 실제 아카이빙 도구가 활용되는 단계들이다.

IIPC에서 제시하는 웹 아카이빙 체인의 각 단계별 내용은 다음과 같다.

2.1 수집 (Acquisition)

아카이빙 과정을 수행하기 위한 가장 첫 번째 단계로서 수집 과정을 통하여 대상 및 범주(target and coverage) 결정, 제한사항(limitations) 검토, 수집 패턴 전략(gathering patterns) 등의 주요 사안들을 결정하게 된다. 수집 패턴 전략이란 통상 장기간에 걸쳐 지속적으로 수행하는 아카이빙 작업에 있어서의 수행 주기에 대한 전략(정기적, 부정기적 등)을 의미한다.

2.2 선정 및 검증 (Focused selection and Verification)

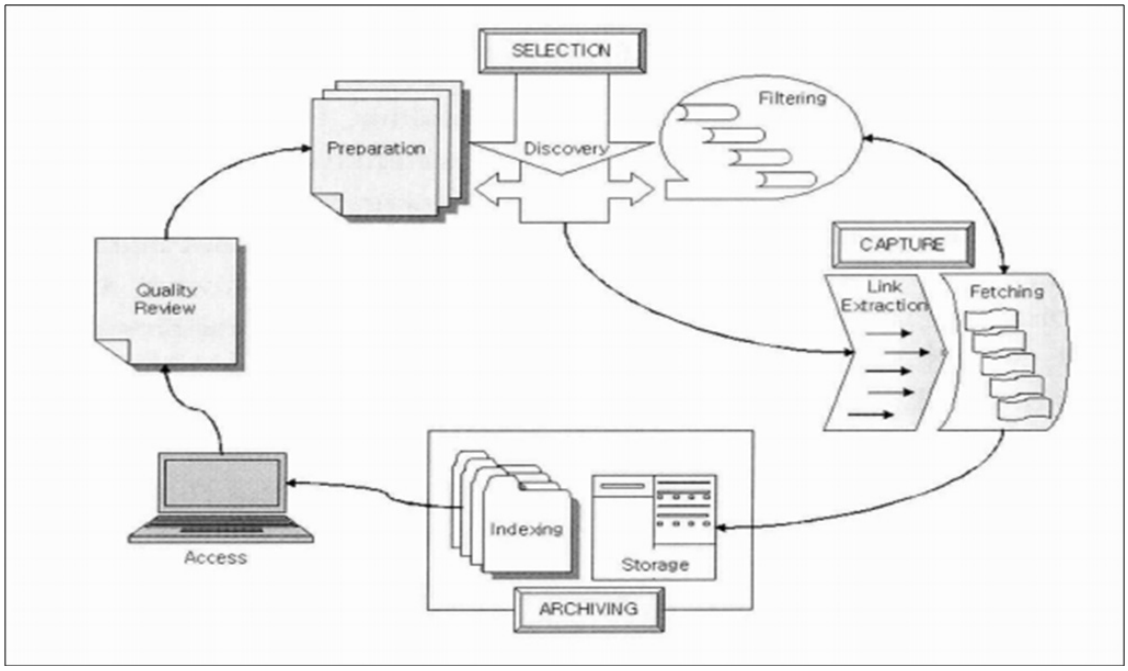
수집 전략에 의하여 아카이브된 웹 자원들을 대상으로 보다 집중적으로 그 적절성을 확인하고 검증하는 단계이다. 대상 웹 자원이 장기보존의 가치가 있는지에 대한 권위 및 신뢰성을 평가한다(appraissal of authority and credibility).

2.3 컬렉션 저장 및 관리 (Collection storage and maintenance)

아카이브된 웹 자원들은 특성과 유형이 다양하다. 컬렉션 저장 및 관리 단계에서는 서로 이질적인 특성을 갖고 있는 웹 자원들을 조정, 제어(manipulation)함으로써, 상호운용성 제고와 콘텐츠 활용을 지원한다.

2.4 접근 및 검색 (Access and finding aids)

아카이브된 웹 자원들을 대상으로 적절한 네비케이션과 URI를 제공하고, 링크 환경을 관리하는 마지막 단계이다. 개선된 브라우징 인터페이스 및 용이한 검색을 위한 작업을 수행한다.



<그림 1> Web Archiving Cycle (Kim and Lee 2007)

3. 단계별 웹 아카이빙 도구 및 특성

앞에서 언급한 웹 아카이빙 과정의 각 단계에 활용할 수 있는 대표적인 웹 아카이빙 도구 유형 및 특성을 정리하면 다음과 같다.

3.1 수집 (Acquisition)

1) Heritrix

Internet Archive와 노르웨이 국가 도서관(Nordic National Libraries)에서 개발한 오픈소스 웹 크롤러이다. 도메인 레벨의 수집 등 포괄적 아카이빙 작업을 수행한다.

2) DeepArc

프랑스 국가도서관에서 개발하였다. portable graphical editor로서 이용자들이 관계형 데이터 모델(relational data model)에서 XLM 스키마로 매핑하는것을 지원하고, 데이터베이스 콘텐츠를 XML문서로 반출(export)하는 것을 지원한다. 심층 웹 문서 수집을 수행한다.

3.2 선정 및 검증 (Focused selection and Verification)

1) Web Curator Tool (WCT)

뉴질랜드 국가도서관과 영국 국가도서관간의 협력에 의하여 시작되었으며, IIPC의 지원으로 개발되었다. 선택적인 웹 하베스팅 절차(selective webharvesting process)를 관리하는 도구로서 이용자 중심의 인터페이스를 제공하고 있다.

2) NetarchiveSuite

덴마크의 국가납본도서관인 The Royal Library와 The State and University Library에서 개발하였다. 기본 기능은 인터넷에서 유용한 자원의 수집을 위한 계획, 스케쥴, 그리고 하베스팅을 운영하는 데에 있다. 좁은 주제 영역의 수집으로부터 국가도메인 규모의 넓은 영역까지 모두 적용할 수 있다.

3.3 컬렉션 저장 및 관리 (Collection storage and maintenance)

1) BAT (BnFArcTools)

프랑스 국가도서관에서 개발하였다. Internet Archive의 ARC, DAT 그리고 CDX 파일 포맷을 처리하는 Perl 패키지이다. 다양한 특성의 파일 제어 및 처리를 지원한다.

3.4 접근 및 검색 (Access and finding aids)

1) Wayback

Internet Archive에서 개발하였으며, 시간의 흐름에 따른 웹 페이지의 다양한 버전들을 아카이빙한다. 이용자들은 원하는 URL과 시간을 입력하여 아카이빙된 웹 페이지에 접근할 수 있다.

2) NutchWAX (Nutch with Web Archive eXensions)

Internet Archive와 노르웨이국가도서관에서 개발하였다. Nutch 검색엔진을 이용하여 웹아카이브의 색인과 탐색을 지원하는 도구이다.

3) WERA (WEb aRchive Access)

Internet Archive와 노르웨이국가도서관에서 개발하였다. 웹 아카이브 탐색 및 네비게이션을 위한 어플리케이션이다.

4) Xinq (XML INQuire)

호주 국가도서관에서 개발하였다. XML기반 콘텐츠들의 탐색과 브라우징을 지원하는 도구이다.

<표 1> 웹 아카이빙 도구 비교

도구	유형	라이센스	URL	적용 과정
Heritrix	Web Crawler	General Public	http://sourceforge.net/projects/archive-crawler	수집
DeepArc	Database archiving tool	General Public	http://sourceforge.net/projects/deeparc/	
WCT	Web Crawler	General Public	http://webcurator.sourceforge.net/	선정 및 검증
Netarchive Suite	Web Crawler	General Public	http://netarchive.dk/suite	
BAT	ARC file manipulation tool	General Public	http://ibnum.bnf.fr/downloads/bat	저장 및 관리
Wayback	Web archive access tool	General Public	http://archive-access.sourceforge.net/projects/wayback/	접근 및 검색
NutchWAX	Web archive search engine	General Public	http://archive-access.sourceforge.net/projects/nutch/	
WERA	Web archive access tool	General Public	http://archive-access.sourceforge.net/projects/wera/	
Xinq	Web archive access tool	Apache Software	http://sourceforge.net/projects/xinq/	

4. 결론

방대한 웹 자원의 생성과 소멸이 급속하게 이루어지고 있는 현대 지식사회에서는 가치있는 웹 자원을 지속적으로 장기보존하는 것이 무엇보다도 중요한 과제이다.

체계적인 웹 아카이빙을 수행하기 위해서는 웹 아카이빙 과정을 이해해야 하며, 각 단계별로 활용할 수 있는 웹 아카이빙 도구의 유형과 특성을 인지하는 것이 필요하다.

본 연구에서는 웹 아카이빙 단계별 활용 도구를 정리하여 제시하였으며, 추후 구체적인 웹 아카이빙 작업 수행시 필요한 도구를 적용하여 활용할 수 있을 것이다.

참고문헌

Heejung K. and Hyewon Lee. 2007. "Development of Metadata Elements for Intensive Web Archiving." *Journal of the Korean Society for Information Management*. 24(2):143-160.

Masanès, J. 2006. *Web Archiving.* Springer.

International Internet Preservation Consortium. [cited 2011. 7. 15.] <//http://www.netpreserve.org/about/index.php>.