웹 아카이빙의 성과와 과제*

Web Archiving: What We Have Done and What We Should Do

서 혜 란(Hye-Ran Suh)**

초 록

이 연구의 목적은 도서관들이 웹 아카이빙이라는 새로운 도전에 대응하여 어떻게 해결책을 모색해 왔으며 앞으로 어떤 과제를 해결해 나가야 할 것인지를 정리하는 것이다. 이 논문에서는 웹 정보자원의 특성을 양적 급성장, 심층 웹의 존재, 웹 정보의 신뢰성에 대한 의문과 역동성, 웹 출판의 무정부성으로 규정하고, 도서관이 왜 웹 아카이빙을 해야 하는가에 대해서 논의하였다. Kurturarw3, PANDORA, Internet Archive를 중심으로 웹 아카이빙 프로젝트의 성과를 검토하였다. 그리고 효과적이고 성공적인 웹 아카이빙을 실현하기 위해서 해결해야 할 정책적 과제와 기술적 과제들을 점검하였다.

ABSTRACT

The purpose of this study is to review what we have done and to identify what we have to do to be successful with Web archiving which is important to preserve our cultural heritage for the next generation. Some characteristics of Web resources as information sources were identified and some difficulties with Web archiving were discussed. The outcome of national and/or international Web archiving projects including Kurturarw3, PANDORA and Internet Archive were reviewed. Policy issues and technological problems of Web archiving we have to solve were listed.

키워드: 웹 아카이빙, 디지털자원, 디지털정보 보존, 디지털 아카이빙
Web Archiving, Digital Resources, Preservation of Digital Information, Digital Archiving

---

* 본 논문은 2004년도 한국비블리아학회 춘계학술발표회에서 발표한 내용을 수정・보완한 것임.

** 신라대학교 정보관리학부 문헌정보학과 교수(hrsuh@silla.ac.kr)

논문접수일자 2004년 5월 31일 논문심사일자 2004년 6월 7일 게재확정일자 2004년 6월 19일

1. 서 론

유럽입자물리학연구소(CERN)에서 과학자 간의 정보공유를 위해 사용되기 시작한 웹은 1990년대 중반에 웹 브라우저가 보급되면서 인터넷을 커뮤니케이션의 중핵도구로 만드는 데 결정적인 기여를 하였다. 지금 인터넷은 온갖 종류의 정보가 유통되는 인류의 보편적 커뮤니케이션 수단으로 자리잡아가고 있다. Lyman & Varian(2003)의 조사 결과에 의하면 2002년 한 해 동안 각종 매체(종이, 필름, 마그네틱매체, 광학매체)에 새로 기록된 정보량은 약 5 엑사바이트(미국의회도서관 장서량의 약 37,000배 정도)이며, 그 가운데 92%는 마그네틱매체(대부분 하드디스크)에 수록되었고 종이에 기록된 것은 0.01%에 불과하다. 같은 기간 동안 전자채널(전화, 라디오, TV, 인터넷)을 통해 전달된 새로운 정보량은 18 엑사바이트이며, 그 대부분(97%)은 전화를 매개로 전달되었지만 약 3%가 인터넷으로 유통되어서 가장 빠르게 성장하는 정보유통채널임이 확인되었다.

World Wide Web Consortium(W3C)(http://www.w3c.org/WWW)은 웹을 "네트워크로 접근할 수 있는 정보의 우주, 인간 지식의 구현"이라고 설명한다. 도서관은 정보 생산자와 정보 소비자 사이에서 매체에 기록된 정보의 유통과정을 관리하고 배포하는 역할과 기록된 정보를 보존하여 후세에 전달하는 책임을 맡은 문화기관이다. 그러므로 웹상에 기록 유통되는 웹 정보자원의 수집과 보존은 도서관의 중요한 관심사가 아닐 수 없다. 다만 웹 정보자원은 그 디지털적 특성으로 인해 인쇄매체를 비롯한 아날로그매체를 중심으로 발전되어 온 종래의 도서관 관행과는 다른 방법론을 요구한다.

이 논문의 목적은 웹 자원의 보존 문제에 대해 우리들이 어떻게 대응하여 해결책을 모색해 왔으며 앞으로 어떤 과제를 해결해나가야 할 것인지를 지금까지 축적된 웹 아카이빙 프로젝트의 경험을 중심으로 정리해보는 것이다. 이를 위해서 제2장에서는 정보원으로서의 웹의 성격과 웹 아카이빙의 필요성을, 제3장에서는 주요한 웹 아카이빙 프로젝트의 성과를 살펴본 후, 제4장에서는 효과적이고 성공적인 웹 아카이빙을 실현하기 위해서 우리에게 남겨져 있는 과제들을 점검해보기로 한다. 한편 이 논문의 리뷰적 성격으로 인해 여기에서는 웹 아카이빙에 관련된 제반 이론을 포괄적으로 다루되 부문별로 구체적으로 분석하지 못하는 한계가 있음을 밝힌다.

2. 정보원으로서의 웹

2. 1 웹 정보자원의 성격

2. 1. 1 양적 성장

웹 공간의 규모와 성장속도에 대한 추정치는 추정방법에 따라 다양하게 발표되고 있지만, 웹의 규모가 대단히 클 뿐 아니라 지속적으로 성장하고 있다는 점을 보여주고 있다는 점에서는 모두 일치한다. 예를 들어서, 1995년 8월부터 매달 인터넷상의 웹 사이트 숫자를 조사해서 발표하고 있는 Netcraft사는 2004년 5월 현재 웹 사이트가 5,000만 개를 넘었다고

발표했다. 이것은 2003년 4월에 4,000만 개를 넘긴지 13개월만의 일인데, 이에 비교해서 웹 사이트가 3,000만 개에서 4,000만 개로 늘어나는 데는 21개월이 걸렸다고 한다(http://news.netcraft.com/archives/web_server_survey.html). 또한 1998년부터 2002년까지 표본조사를 통해 웹 사이트의 특성을 추적한 OCLC의 Web Characterization Project에 의하면 복수의 IP 주소에 중복되어 있는 경우를 제외한 순수한(unique) 웹 사이트 수는 1998년 2,636,000 개에서 2002년 8,712,000 개로 늘어나서 231%의 성장률을 나타냈다(http://wcp.oclc.org).

이처럼 빠르게 늘어나는 웹 사이트를 통해 전달되는 정보량 역시 급속히 늘어나고 있다. Lyman and Varian(2003)은 2002년 현재, 검색엔진에 의해 접근이 가능한 웹의 정보량은 미국의회도서관의 인쇄본 장서량의 17배에 해당하는 167 테라바이트로서 1999년보다 최소한 3배 증가한 것임을 밝혔다.

2. 1. 2 '심층 웹'의 존재

웹의 규모를 보여주는 위 수치들이 웹의 전모를 밝혀주는 것은 아니다. 웹 공간에는 검색엔진을 통해서 자유롭게 접근할 수 있는 '표면 웹(surface Web)' 만이 아니라 일반적 검색엔진에 의해 색인되지 않는 '심층 웹(deep Web)'이 함께 존재하기 때문이다. '숨은 웹(hidden Web)' 또는 '보이지 않는 웹(invisible Web)'이라고도 일컬어지는 '심층 웹'은 대체로 '표면 웹'의 400-550배에 이르는 것으로 추정되고 있다(Bergman 2001).

'심층 웹'은 웹 수집로봇 자체의 기술적 한계에 의해서 생겨나기도 하고, 웹 사이트가 인증과정이나 로봇배제 프로토콜 같은 방법으로 수집로봇의 접근을 거부하여 만들어지기도 한다. 웹 자원의 검색과 아카이빙 측면에서 가장 큰 문제가 되는 것은 전문화된 데이터베이스와 최근 급성장하고 있는 역동적 웹 사이트들이다. 미국 National Oceanic and Atmospheric Administration(NOAA)의 National Climate Data Center, NASA의 Earth Observing System Data and Information System (EOSDIS), UC Berkeley Digital Library, JSTOR, PubMed, Lexis-Nexis, 대형 출판사들의 전자저널 서비스 등은 BrightPlanet사가 조사한 대규모 '심층 웹'의 사례들이다(Bergman 2001).

2. 1. 3 정보의 신뢰성

흔히 웹을 '거대한 도서관'이라고 비유하지만, 실은 개인적 차원의 비공식 커뮤니케이션 수단이나 상업적·오락적 기능을 수행하는 통로로 더 많이 활용되고 있다. 이것은 누구든지 비교적 손쉽게 정보를 웹상에 올릴 수 있다는 사실과 함께 웹을 '거대한 잡동사니' 또는 심지어 '쓰레기장'으로 인식하게 하는 근거가 된다. OCLC의 Web Characterization Project에서는 2002년도의 순수한 웹 사이트 중 36%에 해당하는 3,143,000 개 사이트가 미완성이거나 무의미한 것이라고 확인하였다(http://wcp.oclc.org).

2. 1. 4 역동성

웹 사이트는 빈번하게 생겨났다가 사라지는 역동성을 가지고 있다. 웹 페이지의 평균수명

은 짧게는 44일부터 길게는 2년으로 추정되고 있다(Kenney, et al. 2002). OCLC의 Web Characterization Project에 의하면, 조사 첫 해인 1998년에 확인된 IP 주소의 55%만을 다음 해인 1999년에 찾을 수 있었고 그 비율은 2000년에 35%, 2001년에 25%로 계속 줄어들어서 2002년까지 남아있는 비율은 13%에 불과했다(http://wcp.oclc.org/stats/misc.html). 또 3종의 대표적인 과학저널(New England Journal of Medicine, Science, Nature)에 수록된 과학논문에 각주로 링크된 웹 페이지들 가운데 출판 3개월 후에는 3.8%, 15개월 후에는 10%, 27개월 후에는 13%가 제대로 링크되지 않았다는 조사 결과(Dellavalle, et al. 2003)는 웹 자원의 불안정성을 잘 지적해주고 있다.

더구나 한 웹 사이트의 내용은 지속적으로 수정되고 삭제되기도 한다. 이것은 웹 정보의 최신성을 확보하는 수단이기도 하지만, 웹 아카이빙의 관점에서 보면 큰 문제가 된다.

디지털 자료는 빠르게 발전하는 기술적 인프라에 의존하기 때문에 종이기반 자료에 비해서 미래에 계속 접근할 수 있는 가능성이 훨씬 낮다. 미래의 어떤 시점에서 아카이빙된 웹 정보를 읽을 수 있는 하드웨어와 소프트웨어가 존재하고 작동된다고 보장하기는 쉽지 않기 때문이다.

2. 1. 5 무정부성

웹 자원이 생산되고 유통되는 인터넷은 분산적이고 자율적인 조직이다. 웹 콘텐츠와 그 배포에 대한 결정은 전적으로 웹 사이트 운영자에게 달려있다. W3C 같은 자발적인 협의체가 있지만, 표준의 채택이나 웹 사이트 보존정책을 강제하고 책임지는 조직은 없다. 웹 자원 존재의 증거가 될 수 있는 목록도 없고, 웹의 역동성에도 불구하고 빈번한 변동을 예고하거나 기록하지도 않는다.

2. 2 웹 아카이빙의 필요성

웹은 애초에 학술정보의 공유와 유통 수단으로 만들어졌다. 그 이후 지금까지 학술 및 과학 커뮤니케이션의 수단으로서의 웹의 비중은 지속적으로 확대되고 있다. 전 세계의 연구자들은 웹을 통해서 최신 연구정보를 교환하거나 검색하고, 전자저널 같은 형태로 연구결과물을 배포하고 획득한다(Hendler 2003). 웹이 출현한 후 얼마 지나지 않아서 웹은 연구기관의 범위를 벗어나서 정부, 기업과 상업, 교육, 언론과 출판, 그리고 개인 영역으로 확대되었다.

웹에서 유통되는 정보량이 급증하면서 정보원으로서의 웹에 대한 의존도가 크게 높아졌을 뿐만 아니라 전적으로 웹에만 존재하는 정보자원도 증가하고 있다. 그 가운데는 잘못된 정보를 담고 있거나 아주 일시적인 가치만 가진 것들도 있는 반면에 역사적・문화적・학술적 가치와 법적 증거능력을 가지고 있기 때문에 장기간 보존해야 할 것들도 많이 있다.

그런데 웹의 역동성과 기술의존성으로 인해 웹 자원들이 계속해서 수정되고 바뀌고 삭제되고 있다. 우리가 오늘날 알고 있는 것, 즉 전자적으로 코드화 되고 기록된 것의 대부분이 영원히 사라지게 될 디지털 암흑시대로 옮겨가고 있다는 Kuny(1998)의 경고는 이런 맥락에서 나온 것이다. 우리가 웹 아카이빙을 통해 웹 자원을 수집하고 보존하지 않는다면, 무한한 가치를

가지는 역사적·문화적·학술적 자원들이 미래 세대에게 전달되지 못하게 될 것이다.

3. 웹 아카이빙의 사례

웹 정보의 특성으로 인해 내일의 기억들이 망각 속으로 사라져 가지 않도록 하려면 지금 웹 정보를 수집·보존하는 조치를 취하지 않으면 안 된다는 문제의식이 대두되면서 그에 대한 해결책이 본격적으로 모색되기 시작한 것은 대체로 1990년대 중반 이후라고 볼 수 있다. 최초의 웹 아카이빙 프로젝트는 1994년에 시작된 National Library of Canada의 Electronic Publications Pilot Project (EPPP)이다. 그 때부터 지금까지 웹 자원의 수집과 보존을 위한 시도가 다양한 형태로 이루어지고 있다. 표 1은 그 중 대표적인 사례를 요약한 것이다.

웹 아카이빙 사례들을 웹 아카이빙에 대한 기술적 접근방법에 의해 크게 나누어 보면, 일정한 기준에 의해 수집대상 웹 사이트를 선정한 후 수집하는 선택적 접근방법과 로봇 프로그램을 활용해서 자동적으로 수집하는 포괄적 접근방법이 있다. 한편 추진주체를 보면 국가도서관 또는 국가기록관이 주도하는 경우가 거의 대부분이지만, Internet Archive는 예외적으로 비영리법인에 의해 주도되는 사례이다. 본 논문에서는 선택적 접근방법을 취한 대표적인 사례인 PANDORA와 포괄적 접근방법을 취한 대표적 사례인 Kulturarw3, 그리고 Internet Archive에 대해서 좀더 상세히 설명하도록 한다.

3.1 Kulturarw3(http://www.kb.se/kw3/)

3.1.1 배경

스웨덴의 왕립도서관(Kungl. biblioteket,

(표 1) 주요 웹 아카이빙 사례

국가	사업명	추진주체	수집방법	접근성	규모
호주	PANDORA	National Library of Australia	선택	공개	353 Gb
영국	Britain on the Web (Domain UK)	British Library	선택	비공개	30 Mb
일본	WARP	일본국립국회도서관	선택	공개	524 Gb
미국	MINERVA	Library of Congress	선택	비공개	35 사이트
프랑스	BnF Web archiving initiative	Bibliotheque nationale de France	선택 / 포괄	비공개	1 Tb
스웨덴	Kulturarw3	Kungl. biblioteket(KB)	포괄	제한적 공개	4.5 Tb
핀란드	EVA	Helsinki University Library(National Library of Finland)	포괄	비공개	401 Gb
오스트리아	AOLA	ONB/TU Wien	포괄	비공개	448 Gb
미국	Internet Archive	Internet Archive	포괄	공개	150 Tb

이하 KB)은 1661년에 스웨덴의 인쇄본 출판물을 완벽하게 수집할 책임을 부여받았다. 그 때부터 자국의 문화 및 역사유산을 수집·보존하고 활용시켜 오던 KB는 인터넷에 올려지는 자료의 양이 많아지고 더구나 그 대다수를 인터넷을 통해서만 얻을 수 있게 됨에 따라 오랜 전통을 가진 스스로의 역할을 이어나가기 위해 그 활동범위를 전자출판물로까지 넓혀나가게 되었다. 이에 따라 KB는 1996년에 Kulturarw3 프로젝트를 시작하였다. Kulturarw3의 목적은 온라인으로 접근할 수 있는 스웨덴의 전자문헌을 수집·보존하고 접근·활용시키는 방법을 모색하는 것이다.

3. 1. 2 수집범위와 방법

Kulturarw3는 웹 페이지를 비롯해서 전자저널과 신문, 뉴스그룹과 메일링리스트 등 스웨덴의 온라인 문헌을 파일 포맷에 관계없이 모두 수집한다는 포괄적 수집원칙을 가지고 있다. 이에 따라 일년에 한두 차례 수집로봇 프로그램을 통해 수집범위에 해당되는 모든 자료를 수집한다. 수집로봇은 Lund University Library의 NetLab에서 개발한 공개프로그램인 Combine의 수정판을 사용한다. 1997년 3월 24일부터 같은 해 8월 26일까지 진행된 첫 번째 수집에서 15,700개 웹 사이트로부터 680만 건의 파일, 161 기가바이트를 수집한 이후 수집양은 지속적으로 증가해서 가장 최근인 2002년 8월 7일부터 2003년 6월 4일까지의 열 번째 수집에서는 약 136,800개 웹 사이트로부터 5,690만 건의 파일, 2,195 기가바이트를 수집했다. 현재까지 모두 10차례의 수집을 통해 약 2억 400만 건의 파일, 6,157 기가바이트의 웹 자원이 축적되었다. 아카이브에 등록된 파일포맷의 종류는 약 800여 개로 다양하지만, 텍스트(plain, html, pdf)와 이미지(gif, jpeg)가 대부분(96%)을 차지하고 있다(http://www.kb.se/kw3/ENG/Statistics.htm).

스웨덴의 웹 공간의 전모를 밝혀줄 수 있는 포괄적 수집을 지향하지만 그런 목표가 어느 정도 달성되고 있는지는 불확실하다. 인터넷상에서 스웨덴 웹 자원을 명확히 식별해 내기는 어렵기 때문이다. KB는 ① 스웨덴의 국가 도메인명인 .se를 쓰는 모든 사이트, ② .com, .org, .net 등 국제적인 도메인명이나 .nu[1]을 쓰고 있지만 WHOIS를 활용해서 스웨덴에 소재하고 있는 것으로 식별되는 사이트, ③ 외국에서 만들어진 웹 사이트이지만 스웨덴에 관련된 문제를 다루는 것(스웨덴 여행, 스웨덴문학 작품의 번역 등)을 선택기준으로 한다.

3. 1. 3 저장과 보존

수집된 파일들은 Hierarchical Storage Management(HMS)라는 소프트웨어를 사용하여 Sun 4500 서버에 저장된다. 저장매체는 마그네틱테이프와 약 1.5 테라바이트 용량의 디스크에 저장된다. 수집된 하나의 디지털객체에 관련된 모든 정보(이미지, 사운드 또는 텍스트)가 하나의 파일에 저장된다. 파일은 수집과정에 관련된 메타데이터, 서버에서 디지털객

1) .nu는 남태평양에 위치한 작은 섬나라 Niue의 국가 도메인명이다. Niue는 외국인이나 외국단체에게 자국 도메인명의 사용권을 유료로 판매하고 있는데 특히 스웨덴에서 적극적인 마케팅을 하고 있다. "nu"가 스웨덴어로 "now"를 의미하기 때문이다.

체와 함께 검색된 메타데이터, 디지털객체 자체의 콘텐츠의 세 부분으로 구성된다.

KB는 아카이브 된 파일의 장기보존에서 문제가 되는 것은 저장매체 자체의 수명보다는 소프트웨어와 하드웨어 환경의 단명성이라는 인식 아래 전자문헌의 장기보존을 위한 여러 가지 방법에 대한 실험과 평가를 계속하고 있다. KB는 기본적으로 마이그레이션과 에뮬레이션을 혼용한다는 입장을 취하고 있다(Persson, Arvidson & Mannerhein 2000).

3. 1. 4 이용

처음에 KB는 Kultrarw3 프로젝트에 의해 수집된 아카이브에 대한 공공의 접근을 허용하지 않았다. 이에 대한 법적 규정이 없기 때문이었다. 2002년 5월 8일, 왕립도서관의 인터넷 웹 사이트 수집권한 뿐만 아니라 공공에 대한 제공권을 허용하는 새로운 법이 제정됨에 따라 2003년 6월 16일부터는 구축된 아카이브를 일반인이 이용할 수 있다(http://www.kb.se/Info/Pressmed/Arkiv/2002/020605_eng.htm). 그러나 그 이용은 왕립도서관 내의 터미널에서만 가능하다. 이용자는 해당 웹 페이지를 시간별로 이용할 수 있다.

아카이브의 검색은 웹 브라우저를 통한 일반적인 웹 검색과 마찬가지로 서핑과 검색엔진을 사용한다는 원칙 아래 개발되고 있다. 아카이브 서핑을 위한 소프트웨어의 개발은 상당 정도 진행되었지만 색인에 의한 검색은 아직 실행될 단계가 아니다.

한편 웹 자원의 하베스팅과 아카이빙 분야에서의 경험의 교환과 활동의 조정을 목적으로 결성된 북유럽국가도서관들의 포럼인 Nordic Web Archive(NWA)에서는 2000년 11월부터 웹 아카이브 탐색과 네비게이션용 소프트웨어인 NWA Toolset을 개발하고 있다. 이 프로그램의 개발이 완료되면 스웨덴을 포함한 북유럽국가들에 의해 공동의 웹 아카이브 검색 솔루션으로 활용될 것으로 기대된다.

3. 2 PANDORA(http://pandora.nla.gov.au)

3. 2. 1 배경

PANDORA(Preserving and Accessing Networked Documentary Resources of Australia)는 호주의 온라인 출판물의 장기보존과 활용을 목적으로 National Library of Australia(NLA)의 주도 아래 1996년부터 시작되었다. 1998년 State Library of Victoria의 참여를 필두로 해서 납본권한을 가진 주립도서관들(단 State Library of Tasmania는 PANDORA에 참여하지 않고 Our Digital Island라는 온라인출판물 아카이브를 단독으로 구축하고 있다), 영상과 소리자료를 수집하는 국립기록관인 ScreenSound Australia, Australian War Memorial, Australian Institute of Aboriginal and Torres Strait Islander Studies 같은 문화수집기관들이 파트너로 참여하는 협력체제를 이루고 있다.

3. 2. 2 수집범위와 방법

PANDORA는 선택적 접근법을 채택하고 있다. NLA는 1999년 채택한 'PANDORA 업무처리모델'(http://pandora.nla.gov.au/pandora/bpm.html)에서 그런 결정이 "디지털 출판물을 수집하여 축적하는 일은 복잡하고,

시간이 걸리며, 비용이 많이 드는 일이기 때문에 현 단계에서는 현재 및 미래 시점에서 연구가치가 있을 것으로 생각되는 출판물에 자원을 집중"하기 위한 현실적인 판단에 근거를 두고 있다고 설명하였다. 실제로 NLA는 호주 정부로부터 PANDORA를 위한 별도의 예산을 지원받지 않고 있다.

그렇지만 PANDORA의 선택적 웹 아카이빙은 결코 고육지책이라고 말할 수는 없다. PANDORA의 선택지침(http://www.nla.gov.au/selectionguidelines.html)은 "온라인 출판물에 대해서는 인쇄본에 비해서 더 엄격한 선택이 필요하다"는 이상을 바탕으로 마련된 것이다.

웹 아카이브의 기본적 선택기준은 호주에 관련된 것으로서 그 권위와 장기적 가치가 인정된 인터넷 출판물(웹 사이트와 일부 리스트서브)이며 온라인 포맷만 존재하는 것이다.

일단 대상 웹 사이트가 선정되면 사이트 소유권자와의 합의를 거친 후에 수집을 실시한다. 수집은 수집 소프트웨어를 사용하거나 협의에 의해 웹 출판자가 파일을 물리적 매체에 담아서 보내거나 ftp나 전자우편의 첨부파일로 제출하기도 한다. 수집로봇은 처음에는 University of Colorado에서 개발한 Harvest의 수정판을 사용하다가 지금은 HTTrack과 Teleport Pro를 동시에 사용하고 있다. 복수의 수집 프로그램을 사용함으로써 웹 아카이빙의 성공 확률이 높아진 것으로 평가되고 있다.

캡처 주기와 깊이는 수집대상에 따라 달라진다. 대상 웹 사이트가 너무 큰 경우에는 부분만 수집하기도 한다. 외부 링크는 수집하지 않는다.

2003년 1월 현재 3,300 타이틀(1440만 개 파일, 405 기가바이트)이 축적되었으며 매달 평균 21.5 기가바이트, 약 125 타이틀의 비율로 증가하고 있다. 포맷은 다양하지만 텍스트(HTML)와 이미지(GIF와 JPEG)가 대부분(92%)을 차지한다.

3. 2. 3 저장과 보존

수집된 아카이브는 2001년부터 NLA가 자체 개발한 PANDAS(PANDORA Digital Archiving System)라는 아카이브관리시스템에 의해 관리되고 있다. PANDAS는 PANDORA에 아카이브된 타이틀에 관한 메타데이터를 관리하고, 수집과정 주도, 품질관리, 수집된 자원에 대한 접근관리를 수행한다.

수집된 모든 웹 자원은 MARC에 의해 편목되며, 목록은 National Bibliographic Database와 PANDORA에 참여한 각 파트너들의 자관 목록에 통합된다.

3. 2. 4 이용

아카이브 된 자원에 대한 접근은 PANDORA 웹 사이트를 통해서 할 수 있다. 검색은 주제별 브라우징과 키워드 검색이 모두 가능하다. 대부분은 웹상에서 무료로 이용할 수 있지만 상업적 사이트인 경우에는 일정한 기간 동안 제한된 장소(도서관 열람실)에서만 이용할 수 있도록 제한된다.

3. 3 Internet Archive(http://www.archive.org)

3. 3. 1 배경

Internet Archive는 미국 샌프란시스코에

위치한 비영리단체로서 디지털 형태로 존재하는 역사적 정보자원에 영구적으로 접근할 수 있는 '인터넷 도서관'을 구축한다는 목적을 가지고 1996년에 설립되었다. Internet Archive의 설립을 주도한 Kahle(1997, 2)은 디지털저장에 소요되는 비용이 날로 떨어지기 때문에 웬만한 규모의 컴퓨터 워크스테이션과 데이터저장설비를 갖추기만 하면 소규모의 기술전문가들끼리도 웹 자원을 비롯한 인터넷 정보를 수집하여 영구보존할 수 있다고 주장했다.

3. 3. 2 수집범위와 방법

Internet Archive에 축적되는 데이터들은 Alexa Internet(http://www.alexa.com)을 중심으로 한 파트너들에 의해 수집된다. 주제나 수준 등 수집대상의 범위에 제한을 두지 않고 미래를 위해 광범위하게 수집하는 정책을 유지하고 있다. 1996년에 수집을 시작한 이후, 매달 약 12 테라바이트 정도의 비율로 데이터가 늘어나 2002년 현재 약 150 테라바이트를 축적함으로써 세계 최대 규모의 아카이브 중 하나가 되었다. 이것은 미국의회도서관에 소장된 인쇄본 정보량의 약 7.5배에 해당된다(Kahle 2002).

수집에 사용되는 로봇 프로그램들은 12개월에서 18개월의 주기로 업그레이드된다(Kahle 2002). 수집로봇에 의한 포괄적 수집의 특성으로 인해 로봇배제가 설정된 사이트나 링크가 없는 독립된 사이트, Javascript 등을 사용한 사이트는 수집되지 않는다.

Internet Archive는 특정 주제에 대한 기획 집서를 구축하기도 한다. 웹의 초기 역사에서 중요한 의미를 가지는 사이트들을 모은 'Web Pioneers', Smithsonian Institution과 협력하여 구축한 'Election 1996', Library of Congress와 협력한 'Election 2000 Internet Library', 'September 11 Web Archive' 같은 것이 그 예가 된다.

3. 3. 3 저장과 보존

수집된 데이터는 IDE 하드드라이브가 장착된 일련의 PC들에 저장된다. 웹 데이터는 .arc 파일로 저장되며, 파일에는 27개 필드로 구성된 메타데이터가 붙는다. Internet Archive는 이처럼 분산 저장을 함으로써 불의의 사고로부터 데이터를 보호할 뿐만 아니라, 보존매체인 DLT 테이프의 수명이 30년으로 측정되고 있음에도 불구하고 최소한 10년마다 마이그레이션을 수행함으로써 수집된 데이터를 보존한다는 전략을 가지고 있다. 그와 동시에 소프트웨어의 진화에 대비해서 에뮬레이터의 수집도 계획하고 있다.

3. 3. 4 이용

Internet Archive의 철학은 접근이 허용되지 않는다면 진정한 보존이라고 볼 수 없다는 것이다(Stata 2002). 따라서 Internet Archive는 연구자, 역사가와 학자들을 서비스 대상자로 설정하고, 누구에게나 무료로 접근을 허용하고 있다.

모든 사람들이 좀 더 쉽게 접근할 수 있도록 하기 위한 노력의 일환으로 2001년 10월 24일에 Wayback Machine(http://web.archive.org)이라는 탐색도구가 발표되어 사용되고 있다. 이용자가 웹상에서 보기를 원하는 웹 사이트의 URL을 탐색창에 입력하면 해당 사이트

의 스냅샷들이 일자별로 표시된다. 아직은 색인어를 이용한 검색은 지원되지 않는다. 이런 한계를 극복하기 위한 노력의 결과로서 최근에 원문검색엔진 Recall 시험판이 발표되었다.

4. 웹 아카이빙의 과제

4. 1 기술적 과제

4. 1. 1 아카이빙 범위의 설정

웹 아카이빙의 대상 범위를 설정하는 문제는 크게 보아서 선택적 아카이빙과 포괄적 아카이빙 가운데 어느 접근법이 더 타당한가를 결정하는 것으로 귀결된다.

선택적 아카이빙이란 PANDORA의 사례처럼 미리 정해진 선택기준의 범위에 드는 웹 사이트들만을 대상으로 웹 아카이브를 구축하는 것이다. 이 경우는 사전에 해당 웹 사이트의 저작권자와 수집 및 이용조건에 대한 합의를 이루는 것이 전제가 된다. 선택적 접근법을 취할 경우 얻을 수 있는 장점과 단점은 다음과 같다.

• 장점

① 선택과정을 거치므로 아카이빙 된 모든 웹 자원의 품질이 보장된다.
② 저작권자와의 합의를 거쳐서 공공의 자유로운 이용을 보장할 수 있다.
③ 아카이빙 된 웹 자원 각각에 대해 편목이 가능하므로 국가서지에 통합시킬 수 있다.
④ 인증을 필요로 하는 유료 사이트나 데이터베이스로 구조화된 역동적 사이트들처럼 기술적 이유로 수집로봇이 접근할 수 없는 사이트들도 수집할 수 있다.
⑤ 아카이빙 된 웹 자원들 각각의 장기보존에 필요한 요건을 분석할 수 있고 그에 따른 보존전략을 수립함으로써 영구적 이용을 더 잘 보장할 수 있다.

• 단점

① 어떤 웹 자원이 미래에 필요할 것인지를 지금 결정해야 하기 때문에 일부 중요한 정보가 누락될 위험이 있다.
② 포괄적 접근법에 비해 많은 인력이 필요한 노동집약적 방법이며 인건비가 상승하기 때문에 비용이 많이 든다.
③ 선택적 아카이빙은 대체로 대상 자원의 링크를 포함시키지 않으므로 관련 자원과의 맥락적 의미를 잃어버릴 가능성이 있다.

포괄적 아카이빙이란 Kulturarw3나 Internet Archive의 경우와 같이 주로 수집로봇을 활용해서 일정한 범위 안에 드는 모든 웹 사이트를 광범위하게 수집하는 접근법을 의미한다. 포괄적 접근법을 채택할 경우의 장점과 단점은 다음과 같다.

• 장점

① 미래 세대에게 어떤 정보가 중요할 것인지를 미리 판단하기 어려운 상황에서 미래의 다양한 정보요구에 대응할 수 있다.
② 선택적 접근법을 채택하려면 인건비 때문에 비용이 많이 드는 것에 비해서 컴

퓨터의 데이터 저장비용은 낮아지고 있기 때문에 경제적이다.

• **단점**

① 아카이빙을 위해 점점 더 많은 저장 공간이 필요하게 된다.

② 콘텐츠의 품질 평가가 배제됨으로 인해서 한정된 인적자원과 예산을 수많은 불필요한 정보를 보존하는데 낭비할 우려가 있다.

③ 웹 자원을 중복되게 수집하여 저장할 가능성이 있다.

④ 저작권을 확보하지 못함으로써 일반의 접근을 허용하지 못하게 되면 웹 아카이빙의 의미는 반감될 수밖에 없다.

⑤ 수집로봇의 기술적 한계로 인해 점점 더 증가하고 있는 동적 웹 사이트와 '심층 웹' 정보를 제대로 수집할 수 없다.

두 가지 접근법이 각각 장단점을 가지고 있기 때문에 양자의 장점을 활용하고 한계점을 극복하는 방안으로서 선택적 접근법과 포괄적 접근법을 혼합한 방식을 고려해야 할 필요성이 제기되기도 한다(Day 2003, 30).

법정납본의 대상에 웹 사이트를 포함시킴으로써 푸시 기법으로 웹 아카이빙을 구축하는 방법을 취하게 되면 선택적 수집과 포괄적 수집의 단점들을 상당부분 보완할 수 있을 것이다. 덴마크는 납본법을 개정하여 이미 1998년부터 일정한 기준에 맞는 웹 출판물은 국가도서관에 통지하도록 하는 접근법을 취하고 있으며, 최근에는 영국에서 웹 자료를 납본 대상에 포함시키도록 법을 개정하였다. 그 밖에 호주를 비롯한 여러 나라에서 웹 자원의 법정납본을 목표로 한 노력을 계속하고 있다.

4. 1. 2 수집기술 한계의 극복

표준 html 링크로 연결된 정적인 웹 페이지로 구성된 웹 자원을 수집하는 일은 비교적 쉽다. 그러나 계속 증가하고 있는 javascript나 flash 등 스크립트나 플러그 인 같은 기법을 사용하는 동적인 웹 페이지를 수집하는 일은 간단하지 않다. 스크립트 실행의 결과는 웹 브라우저의 종류 등 많은 사항에 따라 달라지기 때문에 javascript를 채용한 웹 페이지는 성공적으로 처리되기가 어렵다. flash 역시 플러그인을 사용할 뿐만 아니라 상용포맷이기 때문에 그런 웹 페이지를 수집하기는 매우 어렵다.

웹 사이트의 운영자가 로봇배제표준을 사용하여 웹 사이트에 robots.txt 파일(http://www.robotstxt.org/wc/exclusion.html)을 심어놓은 경우에도 로봇에 의한 수집이 불가능하다. 데이터베이스를 기반으로 하는 웹 사이트도 검색엔진과 마찬가지로 대부분 수집로봇을 통해서 아카이빙이 불가능하다. 이처럼 '심층 웹'의 존재는 특히 수집로봇에 의존하는 웹 아카이빙에 심각한 문제를 제기하고 있다. 최근 Google이나 Yahoo 같은 검색엔진업체들과 컴퓨터 연구자들은 검색의 관점에서 이 문제의 해결에 투자하고 있지만(Olsen 2004), 이 문제는 웹 아카이빙의 관점에서도 많은 연구가 필요하다(Day 2003, 16)

4. 1. 3 수집주기와 수집시간

'스냅샷'에 의한 아카이빙에서 문제가 되는 것은 각 스냅샷 간의 간격보다 수명이 짧은 데

이터들을 너무 많이 잃어버리게 된다는 점이다. 이 문제를 해결하기 위해서는 수집로봇으로 하여금 웹 자원의 변동 주기를 자동적으로 체크하도록 해서 그에 따라 수집주기를 조절할 수 있을 것이다(Arvidson, Persson, & Mannerheim 2001).

진정한 '스냅샷'이 되기 위해서는 수집에 걸리는 시간이 가능한 한 짧아야 한다. 그러나 Kulturarw3의 경험에서 보듯이 실제로는 수집 대상이 되는 웹 사이트가 방대하기 때문에 완전히 수집하는데 상당한 시간이 걸린다. 그러다보면 한 페이지가 서로 다른 시간에 수집된 몇 가지 객체로 이루어지기 때문에 결과적으로 실제로는 존재하지 않았던 웹 페이지가 아카이빙 됨으로써 진본성(authenticity)의 훼손을 가져올 수 있다(Howell 2003).

4. 1. 4 장기보존

아카이빙은 "장기적 가치를 가진 디지털자원을 저장하고 접근하고 관리할 수 있는 안전한 장소에 가져다 놓는 것"이고, 장기보존(long-term preservation)이란 "아카이빙이 이루어진 출판물들이 장기적으로 계속 생존하면서 접근이 가능하도록 보장하는 전략을 제안하고 활동을 수행하는 과정"이다(Webb 2000). 웹 자원의 장기적 이용을 보장하려면 아카이빙과 함께 장기보존을 위한 조치가 필수적이다. 따라서 웹 자원의 아카이빙 활동과 그것을 통해 축적된 아카이브의 관리시스템 운영에는 반드시 장기보존의 요소가 중요하게 고려되어야 할 것이다.

디지털자료의 진본성과 신뢰성 보장이 증명된 장기보존 기법은 아직 존재하지 않는다. 지금까지 기술보존(technology preservation), 기술 에뮤레이션(technology emulation), 정보 마이그레이션(information migration), 인캡슐레이션(encapsulation), 디지털 타블렛(Digital Tablet) 같은 여러 가지 기법들이 제시되었으며, 이 기법들은 각각 장단점을 가지고 있는 것으로 보고되고 있다(Lee, et al. 2002). 이 문제의 해결 역시 좀더 시간이 필요한 것으로 보인다.

웹 아카이빙의 장기보존과 관련해서 주목해야 하는 것은 2002년 1월에 ISO 표준(ISO 14721:2002)으로 확정 공포된 OAIS 참조모형(Reference Model for the Open Archival Information System)이다. 디지털 정보의 '장기보존'과 '접근제공'이라는 두 가지 기본기능의 충족을 위한 개념적 틀이 만들어진 것이다(Lavoie 2004). 앞으로 남은 과제는 OAIS 참조모형에 기반을 둔 디지털보존시스템을 구현해내는 일이며, 여기에는 시스템 아키텍처의 설계, 저장 및 처리기술, 데이터베이스 설계, 보존 메타데이터 등 수많은 과제가 포함되어야 한다.

4. 1. 5 이용가능성의 보장

도서관이 웹 자원의 단순한 저장소가 아니므로 아카이브 된 웹 자원을 효율적으로 검색할 수 있도록 하는데 많은 관심을 기울여야 한다. 이 문제는 저작권과 관련된 법률적 과제이기도 하지만, 기술적 측면에서의 개선도 수반되어야 한다. 메타데이터의 자동 생성과 대규모 웹 아카이브의 자동색인과 자동편목 기법이 만들어진다면 아카이브에 접근하는 새로운 방식으로 추가될 수 있을 것이다.

역동성이 강한 웹 자원의 이용가능성의 제

고와 관련해서는 DOI 등 항구적 식별자(persistent identifier)의 적용, 변경된 URL의 자동추적 기술의 적용도 반드시 고려되어야 한다.

4. 2 정책적 과제

4. 2. 1 협력체제의 구축

협력은 웹 아카이빙의 성공적 추진을 위한 핵심 키워드가 되고 있다.

웹은 지구 전체를 포괄하는 현상이다. Internet Archive 같은 예외가 있기는 하지만, 어떤 웹 아카이빙 프로젝트라도 단독으로 웹 전체를 대상으로 수집·보존하려는 계획을 세우기는 어렵다. 이미 유럽이나 북미의 정보선진국들 간에는 국제적 차원이나 지역적 차원의 협력을 통해 각종 표준이나 지침의 개발을 함께 하거나 경험을 공유하는 경향을 보이고 있다. 그런 경향이 나타나는 것은 첫째, 자원과 인력의 통합을 통한 발전의 촉진, 둘째, 데이터의 수집과 저장 등에 규모의 경제를 실현함으로써 각 국가의 비용부담 절감, 셋째, 국가간 공동활용체제 구축 같은 장점을 추구하기 위한 노력 때문이다. International Internet Preservation Consortium(http://netpreserve.org), NEDLIB Harverster Project(http://www.csc.fi/sovellus/nedlib), Nordic Web Archive(http://nwa.nb.no) 등이 그 사례들이다.

같은 맥락에서, 한 국가 내에서의 협력체제 구축도 필요하며 또 그런 시도들이 많이 이루어지고 있다. 단일 기관이 국가적 웹 아카이빙을 모두 떠맡기는 어렵기 때문이다. 국가적 차원의 웹 아카이브를 비용효과적으로 구축하기 위해서는 국가도서관과 기록관을 비롯해서 다양한 관련 기관들이 긴밀한 협력 체제를 구성할 필요가 있다. 특히 웹 아카이빙에는 콘텐츠 생산자, 출판자, 배포자, 이용자 등 다양한 입장의 이해당사자들이 관련되기 때문에 그들의 의견을 통합 조정할 수 있는 리더십의 존재가 중요하다.

4. 2. 2 법률적 조치

웹 아카이빙에는 저작권, 납본, 개인정보 보호, 콘텐츠의 진본성과 신뢰성을 보장하는 보존 책임 등의 법률적 문제가 관련되어 있다. 웹 자원을 비롯한 디지털 정보의 수집, 저장, 보존, 유통과 활용에 영향을 미치는 법률적 장치들은 전통적 매체와는 다른 특성을 가지고 있는 디지털 정보의 속성으로 인해서 이해당사자들 간에 새로운 논쟁거리가 되고 있다. 이해당사자들에는 정보의 생산자, 소유자, 웹 아카이브 관리자, 공공정책 당국자, 그리고 정보이용자들이 포함된다. 법률적 문제는 디지털 정보 보존의 필요성에 대한 이해당사자간 선의의 합의 아래에서 신중하면서도 동시에 적극적으로 추진되어야 할 과제이다.

4. 2. 3 재정

웹 아카이빙의 성과는 많은 자원의 투자에 비해서 볼 때 단기간 내에 가시적으로 성과가 드러나는 일이 결코 아니다. 그러므로 웹 아카이빙의 성공을 위해서는 지속적이고 안정적인 재원을 확보하는 일이 중요하다.

그러나 PANDORA나 Kulturarw3를 비롯해서 전 세계적으로 진행되는 웹 아카이빙

관련 프로젝트들 중에 별도의 재정지원을 충분히 받는 사례는 의외로 찾아보기 어렵다. 그런 점에서 의회로부터 1억 달러의 자금지원 승인을 얻어낸 미국 Library of Congress의 NDIIPP는 상당히 예외적인 사례로서 주목하지 않을 수 없다.

웹 아카이빙의 공공성으로 볼 때 그 재정은 기본적으로 공적자금에 의해 충당되어야 하지만, 필요한 자금의 규모로 볼 때 민간자금의 지원이 절대로 필요하다. 그러나 정부는 물론 민간부문에서도 디지털자원의 보존에 대한 인식 수준은 아직 그리 높지 않은 것도 사실이다. 이 부문에 대한 지속적인 홍보와 설득이 이루어질 필요가 있다.

4. 2. 4 기타

웹 아카이빙을 위해서는 다양한 관련 기술을 가진 기술 인력과 함께 그들의 공동 활동을 조정할 능력을 가진 리더가 필요하다. 필요한 인력의 경력개발과 교육훈련, 효율적 조직을 통해 이들의 시너지 효과를 발휘할 수 있는 정책이 마련되는 일이 중요하다.

또한 관련 기술의 개발이나 표준화를 위해서는 아직 가야 할 길이 멀기 때문에 연구개발에 대한 지속적인 투자와 관심, 국제적 협력사업에의 적극적 참여가 이루어져야 한다.

5. 결 론

웹 아카이빙에 관한 이론적 연구와 실천적 경험의 축적이 아직 성숙단계에 들어섰다고 보기는 어렵더라도 적어도 걸음마단계는 넘어섰다고 볼 수 있다. 따라서 지금 디지털 문화유산이 사라져버릴 위험을 걱정하고 웹 자원 수집 보존의 필요성을 주장하는 것은 다소 새삼스럽고 진부하게 들릴지도 모른다. 그럼에도 불구하고 그것을 다시 되뇌는 이유는 그 사안이 국가적 또는 국제적 차원에서의 치밀한 계획과 자원투자 속에서 다양한 이해당사자들의 긴밀하고 우호적인 협력을 통해서만 성공을 기대할 수 있는 것이라고 보기 때문이다.

웹 아카이빙 문제에는 많은 이해당사자들이 복잡하게 관련되어 있다. 아직까지는 표준이나 합의된 해결책도 별로 없다. 기술적 문제들이 많긴 하지만 기술적 문제들이 전부는 아니다(Warner 2002, 47).

전통적으로 종이기반 정보자료를 수집하고 이용시키고 보존하는 일은 도서관에게 일임되어 왔다. 이는 도서관이 자료의 소유와 이용제공을 통해 상업적 또는 정치적 이익을 추구하지 않을 것이며, 일단 도서관에 들어온 종이자료는 장기보존 될 것이라는 인식이 있기 때문이다. 다시 말하면 종이기반 자료의 경우 도서관은 '신뢰받는 보관소(trusted repositories)'인 것이다(Howell 2003). 웹 자원의 경우에도 마찬가지로 그것이 신뢰받는 보관소에 있다는 사실이 다른 모든 활동을 평가하는 기초가 될 것이다(Hirtle 2000, 20).

대규모 디지털 문화유산의 신뢰받는 보존소의 성격과 책임을 규정하기 위해 RLG와 OCLC가 공동으로 구성한 워킹그룹은 디지털 자료의 신뢰받는 보관소를 "관리하는 디지털 자원을 현재 및 미래 시점에서 지정된 이용자 집단에게 신뢰할 수 있는 장기적 접근을 제공하는 것을 사명으로 하는 기관"(RLG/OCLC

Working Group 2002, 5)으로 정의하고 OAIS 참조모형의 준수, 행정의 책임성, 조직의 지속성, 재정능력, 기술과 절차의 적절성, 시스템 안전성, 절차의 투명성을 그 핵심 속성으로 열거하였다. 그리고 국제적 및 국가적 차원에서 디지털 자료의 장기적 보존을 책임져야 할 기관들에게 디지털보존소의 인증절차 마련, 반드시 보존되어야 할 디지털 자료의 주요 속성을 확인하는 도구에 대한 연구와 개발, 협력보존 네트워크와 서비스 모델에 대한 연구와 개발, 장기보존을 명백하게 지원하는 디지털 객체의 확인 시스템 개발, 디지털 보존과 지적재산권 간의 복잡한 관계에 관한 조사와 정보배포, 지속적 접근을 가장 잘 보장해주는 기술적 전략의 결정, 장기적 관리에 필수적인 최소수준의 메타데이터 설정과 그것을 자동적으로 생성·추출하는 도구의 개발을 수행하도록 권고하였다.

인터넷의 보급과 활용 면에서 세계적 선도그룹에 속하는 우리나라로서는 당연히 웹 아카이빙에 대해서 관심을 기울이고 웹 자원의 장기적 보존을 위해 상당한 노력을 기울여야 할 것이다. 그럼에도 불구하고 인터넷 강국이라는 명성에 비해서는 웹 자원의 보존에 대한 관심과 투자가 저조했다고 볼 수 있다. 다행히 최근 국립중앙도서관은 국가디지털도서관 건립사업과 함께 현세대의 중요하고 가치 있는 온라인 디지털자원을 후세대에 물려줄 수 있는 기반환경 조성을 위해 필요한 기초적 활동을 시작하는 단계에 있는 것으로 알려져 있다. 한편 민간부문에서는 6개 시민단체들(다음세대재단, 문화연대, 사이버문화연구소, 정보공유연대, 진보네트워크센터, 함께하는 시민행동)이 함께 추진하는 '정보트러스트운동'(http://www.infotrust.or.kr)을 추진하고 있다. 2003년부터 시작된 이 운동은 "인터넷상에서 사라져가고 있는 디지털 정보를 복원하고, 보전할 가치가 있는 사이버공간의 지식과 정보를 시민들의 자발적 참여와 모금으로 공공화 하는 운동"을 표방하고, 인터넷 역사의 정리, 인터넷 공간에서 사라져가는 디지털 정보의 복원과 가치있는 정보의 보전 활동을 계획하고 있다.

지금까지 주요 국가의 국가도서관이 주도하는 프로젝트들을 중심으로 해서 상당한 이론과 실천 경험들이 축적되어 왔지만 성공적인 웹 아카이빙을 위해서는 아직까지 수많은 정책적 과제와 기술적 과제들이 해결을 기다리고 있다. 그 과제들은 결코 어떤 단일 국가나 기관에 의해서 완전하게 풀릴 수 있는 것은 아니라고 본다. 따라서 우리나라도 대내적으로는 웹 아카이빙에 관련된 모든 기관들과 이해당사자들의 참여 속에서 국가적 차원의 웹 아카이빙 네트워크를 구축하고, 대외적으로는 선진국의 경험을 분석하고 벤치마킹하는 한편 축적된 지식을 국제적으로 공유할 수 있도록 장기적인 관점의 로드맵을 작성해야 할 것이다. 그리고 그 과정에서 국립중앙도서관의 리더십이 중요한 역할을 해야 할 것이라고 생각한다.

참 고 문 헌

Arvidson, Allan. 2002. The Collection of Swedish Web Pages at the Royal Library: the Web Heritage of Sweden. IFLA Council and General Conference, 68th, Glasgow, August 18-24, 2002. http://www.ifla.org/IV/ifla68/papers/111-163e.pdf

Arvidson, Allan, Krister Persson and Johan Mannerheim. 2001. The Royal Swedish Web Archive: a 'Complete' Collection of Web Pages. *International Preservation News* 26: 10-12.

Bergman, Michael K. 2001. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing* 7(1) http://www.press.umich.edu/jep/07-01/bergman.html

Charlesworth, Andrew. 2003. *Legal Issues Relating to the Archiving of Internet Resources in the UK, EU, USA and Australia: a Study Undertaken for the JICS and Wellcome Trust.* http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf

Day, Michael. 2003. *Collecting and Preserving the World Wide Web: a Feasibility Study Undertaken for the JICS and Wellcome Trust.* http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

Dellavalle, Robert P. et al. 2003. Going, Going, Gone: Lost Internet References. *Science* 302(5646): 787-788.

Hendler, J. 2003. Science and the Semantic Web. *Science* 299(5606): 520-521.

Hirtle, P. B. 2000. Archival Authenticity in a Digital Age. In *Authenticity in a Digital Environment.* Washington, D.C.: Council on Library and Information Resources, 8-23. http://www.clir.org/pubs/abstract/pub92abst.html

Howell, Allan G. 2003. *Preserving Digital Information: Challenges and Solutions: Workbook.* http://www.alanhowell.com.au/papers/pdi_wkb.pdf

Kahle, B. 1997. Preserving the Internet. *Scientific American* 276(3): 72-73.

Khale, B. 2002. Editors' Interview: the Internet Archive. *RLG DigiNews* 6(3) http://www.rlg.org/preserv/diginews/diginews6-3.html#interview

Kuny, Terry. 1998. The Digital Dark Ages?: Challenges in the Preser-

vation of Electronic Information. *International Preservation News* 17: 8-13. http://www.ifla.org/VI/4/news/17-98.htm#2

Lavoie, Brian F. 2004. *The Open Archive Information System Reference Model: Introductory Guide.* OCLC Online Computer Library Center, Inc. and Digital Preservation Coalition.

Law, Cliff. 2001. PANDORA: the Australian Electronic Heritage in a Box. *International Preservation News* 26: 13-17.

Lee, Kyung-Ho, et al. 2002. The State of the Art and Practice in Digital Preservation. *Journal of Research of the National Institute of Standards and Technology* 107(1): 93-106.

Lyman, Peter. 2002. Archiving the World Wide Web. In *Preserving Our Digital History: Plan for the National Digital Information Infrastructure and Preservation Program.* Washington, D.C.: Library of Congress.

Lyman, Peter and Hal R. Varian. 2000. How Much Information? 2000. http://www.sims.berkeley.edu/research/projects/how-much-info/

Lyman, Peter and Hal R. Varian. 2003. How Much Information? 2003. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Mannerheim, Johan. 2001. The New Preservation Tasks of the Library Community. *International Preservation News* 26: 5-9.

Olsen, Stefanie. 2004. Yahoo crawls deep into the Web. http://zdnet.com.com/2100-1104_2-5167931.html

O'Neill, Edward T., Brian F. Lavoie, Rick Bennett. 2003. Trends in the Evolution of the Public Web: 1998-2002. *D-Lib Magazine* 9(4) http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

Persson, Krister, Allan Arvidson, Johan Mannerheim. 2000. The Kulturarw3 Project - The Royal Swedish Web Archiw3e. http://www.kb.se/kw3/articles/article000605.pdf

Phillips, Margaret E. 1999. Ensuring Long-Term Access to Online Publications. *Journal of Electronic Publishing* 4(4) http://www.press.umich.edu/jep/04-04/phillips.html

RLG/OCLC Working Group on Digital Archive Attributes. 2002. *Trusted Digital Repositories: Attributes and Responsibilities.* Mountain View, Calif.: Research Libraries

Group. http://www.rlg.org/longterm/repositories.pdf

Stata, Raymie. 2002. Presentation of the Internet Archive. ECDL Workshop on Web Archiving, 2nd., Rome, Italy, September 19, 2002. http://bibnum.bnf.fr/ecdl/2002/

Warner, Dorothy. 2002. Why Do We Need This in Print? It's on the Web...: a Review of Electronic Archiving Issues and Problems. *Progressive Librarian* 19/20: 47-64.

Webb. Colin. 2000. Towards a Preserved National Collection of Selected Australian Digital Publications. in *Preservation 2000: an International Conference on the Preservation and Long Term Accessibility of Digital Materials, 7/8 December 2000, York, England, Conference Papers.* http://www.rlg.org/events/pres-2000/webb.html