국립중앙도서관

ISSUE PAPER

이슈페이퍼

웹자원 아카이빙(OASIS) 현황 및 사례, 미래 발전방안

Vol. 14
2023. 3.

문화체육관광부
국립중앙도서관

웹자원 아카이빙(OASIS) 현황 및 사례, 미래 발전방안

안경자 주무관
(국립중앙도서관 온라인자료과)

Vol. 14
2023. 3.

국립중앙도서관
ISSUE PAPER
이슈페이퍼

국립중앙도서관 **이슈페이퍼 제14호**

발행일 2023년 3월 30일
발행처 국립중앙도서관
발행인 국립중앙도서관장
주소 서울특별시 서초구 반포대로 201
전화 02-590-0578
팩스 02-590-0546
누리집 https://nl.go.kr
ISSN 2765-3005

| 주요 키워드 |
웹자원 아카이빙, 웹 아카이빙, OASIS, 오아시스, 인터넷 보존

웹자원 아카이빙(OASIS) 현황 및 사례, 미래 발전방안

안경자 주무관
(국립중앙도서관 온라인자료과)

요약

국립중앙도서관은 2004년부터 소멸되기 쉬운 웹자원을 국가 지식문화유산으로 수집·보존하는 오아시스 사업(Online Archiving & Searching Internet Sources, OASIS)을 추진하고 있다. 오아시스는 국가 도메인 웹사이트와 웹에서 유통되는 자료를 수집하는 대규모 사업이다. 국립중앙도서관은 국내에서 유일하게 국가 단위의 웹사이트를 포괄적으로 수집하고 있으며, 특히 국가 재난 및 행사 등 중요한 주제를 대상으로 웹 컬렉션을 구축하여 오아시스 누리집을 통해 제공하고 있다.

2000년대 인터넷 발달과 더불어 대량의 온라인 정보자원이 유통되기 시작면서 웹은 가장 편리한 지식정보 채널로 자리 잡기 시작했다. 2003년 10월 유네스코에서 "도서관을 포함한 유산기관(heritage institution)은 디지털 형태로 생산·배포·액세스·유지되는 디지털 정보자원을 보존해야 하며 보편적 접근용이성을 보장해야 한다"는 디지털 유산의 보존에 관한 헌장(유네스코한국위원회, 2003)이 공표되었으며, 이를 계기로 세계 주요 국가도서관을 주축으로 웹 아카이빙에 대한 고민과 시도가 시작되었다. 2003년 인터넷 자원의 수집 및 보존을 위한 전 세계적인 협력과 공조를 위해 국제인터넷보존컨소시엄(International Internet Preservation Consortium, IIPC)이 설립되었으며, 웹 아카이빙을 위한 국제 표준과 기술을 개발하여 제공하고 인터넷 보존 이니셔티브 활동을 진행하고 있다. 하지만 양적으로 방대할 뿐만 아니라 수시로 갱신되고 변화하는 동적 자원인 웹자원을 수집하는 것은 상당히 어려운 작업이다. 웹자원은 수집 및 보존, 이용에 한계가 있으며, 무결성 보장이 어렵고 진본성을 판별해야 하는 등의 여러가지 문제를 안고 있다.

이러한 과제 속에서 디지털 자원 및 문화 보존을 위한 체계적인 웹 아카이빙이 요구되며, 전 세계 도서관은 다자간 협업과 창의적인 실험을 통해 이를 성공적으로 해결하기 위해 노력하고 있다.

국립중앙도서관은 어려운 여건 속에서도 미래 정보자원인 웹자원을 사라지기 전에 수집 및 보존하기 위해 노력하고 있으며, 오아시스 개선과 활성화를 위한 다양한 시도를 진행하고 있다. 오아시스가 직면하고 있는 당면 과제를 살펴보고 미래를 위한 발전방안을 제안하고자 한다.

주요 키워드 웹자원 아카이빙, 웹 아카이빙, OASIS, 오아시스, 인터넷 보존

I. 웹자원 아카이빙(OASIS) 현황

1. 오아시스 개요

국립중앙도서관은 『도서관법』 제22조(온라인 자료의 수집)에 의거하여, 소멸되기 쉬운 공개 인터넷자원(웹사이트 및 웹자료)을 국가 지식문화유산으로 수집·보존하고 후대에 전승하기 위해 2004년부터 오아시스 사업(Online Archiving & Searching Internet Sources, OASIS)을 추진하고 있다. 오아시스는 대규모 국가 웹자원을 수집 및 보존하는 국내 유일의 사업이다. 수집 대상 자료는 인터넷 웹사이트와 웹문서, 동영상, 이미지 등 웹사이트에서 공유되고 있는 웹자료이다.

2. 추진경과

디지털 정보자원의 영구보존을 꿈꾸다! OASIS

2000년대 인터넷 발달과 더불어 인터넷 정보자원에 대한 수집 및 보존의 필요성이 대두되었다. 2003년 10월 유네스코에서 "도서관을 포함한 유산 기관(heritage institution)은 디지털 형태로 생산·배포·액세스·유지되는 디지털 정보자원을 보존해야 하며 보편적 접근용이성을 보장해야 한다"는 디지털 유산의 보존에 관한 헌장(유네스코한국위원회, 2003)이 공표되었다. 이를 계기로 국립중앙도서관은 2003년 온라인 디지털 정보자원 수집 및 보존을 위한 전담팀을 구성하였으며, 2004년부터 웹사이트 시범 수집을 시작하였다. 2006년에는 오아시스 누리집을 구축하여 대국민 웹사이트 검색 및 이용 서비스를 개시하였다. 2008년에는 국제인터넷보존컨소시엄(International Internet Preservation Consortium, IIPC)에 가입하여 웹 아카이빙을 위한 전 세계 도서관 컨소시엄에 참여하기 시작하였다.

표준 보존파일(WARC)부터 포괄적 수집까지

오아시스는 2011년부터 IIPC에서 권장하는 헤리트릭스(Heritrix) 웹자원 수집기[1](이하 수집기)와 국제 표준 웹사이트 보존 파일 형식인 WARC[2] 파일형식을 채택하여 현재까지 적용하고 있다. 지속적으로

1 2003년 인터넷 아카이브사와 노르웨이국립도서관이 함께 개발한 웹사이트 수집을 위한 오픈 소스 웹 크롤러로 현재 대부분의 웹 콘텐츠를 수집하는 도구로 사용되고 있음

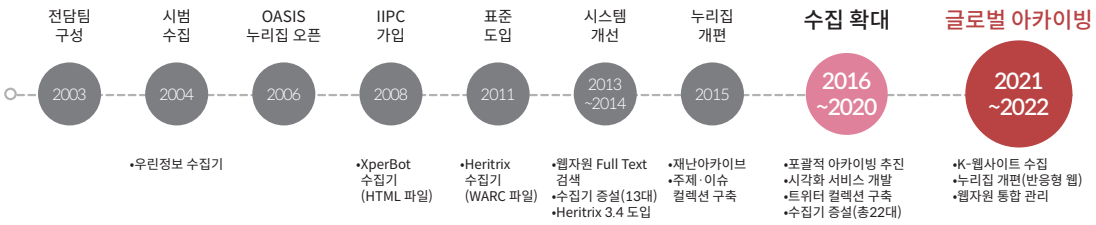
2 웹사이트 수집 시 저장되는 보존용 파일 포맷으로 ISO 28500 표준

수집기를 증설한 결과 2022년 말 현재 총 22대 수집기를 가동하고 있다. 2016년에는 주요 웹사이트를 선별적으로 수집해오던 기존 방식에 더해 국가 도메인 웹사이트(약 110만 건) 전체를 대상으로 한 포괄적 수집을 병행하기 시작하면서 웹사이트 수집을 본격적으로 강화하였다. 또한 소셜미디어 급증에 따라 최초의 소셜미디어 컬렉션인 트위터 컬렉션을 자체 개발을 통해 구축하였으며, 웹 데이터 분석 및 시각화 서비스인 태그 클라우드, 웹 트렌드 서비스를 개발하였다.

웹자원 큐레이션의 시작, 국가재난아카이브

2015년부터는 국가 재난 관련 웹자원을 집중적으로 수집·보존하기 위해 국가재난아카이브[3]를 구축하기 시작하였다. 이와 더불어 선거, 올림픽 등 국가 중요 행사를 중심으로 주제 및 이슈 관련 웹자원 컬렉션을 구축하는 등 누리집 개편을 통해 웹자원 큐레이션 서비스를 강화하였다.

2020년에는 인류사를 바꾸어놓은 전대미문의 재난인 코로나19 관련 웹 기록을 발생 시점부터 집중 수집하여 현재까지 코로나19 재난아카이브[4]를 구축·운영하고 있다. 2022년에는 반응형 웹을 지원하도록 오아시스 누리집을 최신 흐름에 맞게 개편하였으며, 한류 확산과 더불어 해외에서 생산·제작되고 있는 K-웹사이트에 대한 조사 및 수집을 시작하여 글로벌 K-웹자원 아카이빙을 확대하고 있다.



[그림 1] 오아시스 추진경과

3. 수집·구축·서비스 현황

국립중앙도서관은 웹자원 수집 시 선택적 수집 방법과 포괄적 수집 방법을 병행하고 있다. 먼저 선택적 수집 방법은 ① 정부·공공기관, 언론기관, 상업기관, 문화예술기관 등 신설 또는 개편된 주요 기관의 개별

3 코로나19를 비롯하여 세월호사고, 메르스, 아프리카돼지열병 등 주요 국가재난 40건에 대한 총 8만 건의 재난 기록을 보존('23.2.기준)

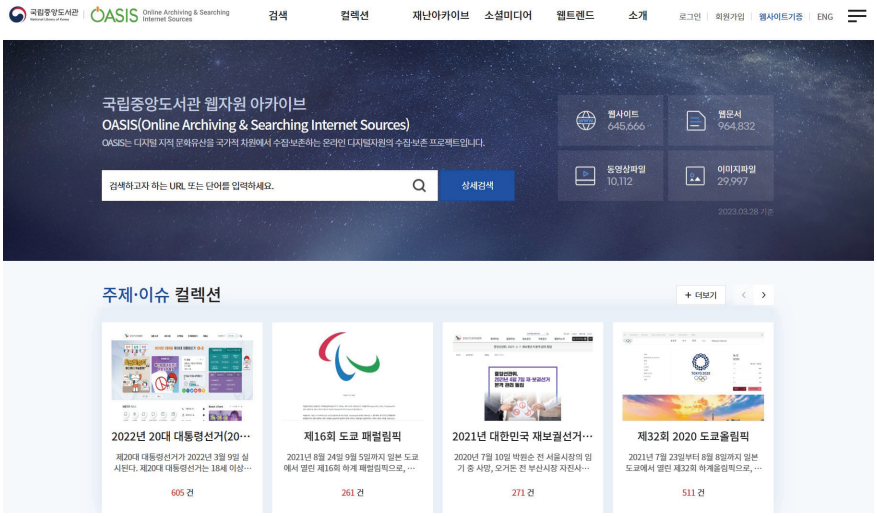
4 코로나19 발생('20.1.20.)시점부터 관련 웹자원 16,987건을 수집 및 보존('23.2.기준)

웹사이트, ② 국가적 재난 및 행사, 특정 주제와 관련된 웹자원(웹사이트, 웹자료)을 집중 수집하는 방식이다. 포괄적 수집 방법은 매년 한국인터넷진흥원(KISA)에 등록되어 있는 대한민국 국가 도메인(KR) 웹사이트 전체(약 110만 건)를 A부터 Z까지 빠짐없이 모두 수집하는 방식이다.

수집부터 메타데이터 작성까지

온라인 자료 수집지침(국립중앙도서관 온라인자료과, 2022)에 따라 웹자원의 가치성, 신뢰성, 유일성 등을 검증하고 정상 수집이 확인된 경우에 한하여 국가 장서로 등록하여 영구 보존한다. 선택적으로 수집된 웹사이트의 경우 일반 장서와 같이 메타데이터(MODS)[5]를 구축하여 활용되고 있으나, 포괄적으로 수집한 웹사이트의 경우 예산 부족으로 인해 대부분 메타데이터가 미구축된 상태이다. 정상 수집 후 등록된 웹자원은 『저작권법』 제31조(도서관등에서의 복제 등)에 따라 국립중앙도서관 내에서 이용 가능하며, 저작권자가 허락한 경우에 한해 외부 이용자료로 제공되고 있다.

2022년 말까지 웹사이트 101만 건과 웹자료 144만 건, 총 245만 건의 웹자원(1,007TB)을 수집하였으며, 코로나19를 비롯하여 선거, 올림픽 등 국가 재난·행사 및 주제·이슈 관련 컬렉션 총 154개를 구축하여 오아시스 누리집(https://nl.go.kr/oasis)을 통해 제공하고 있다.



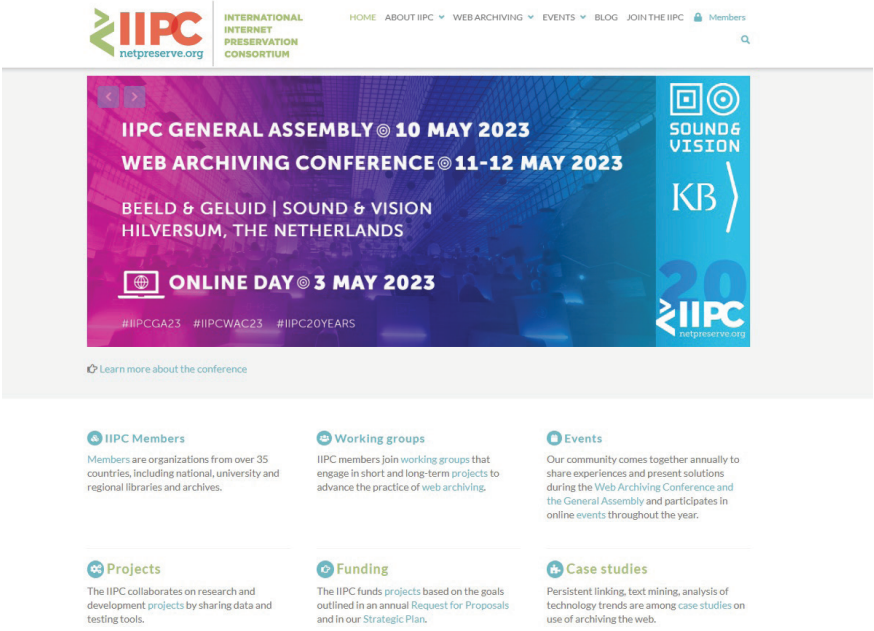
[그림 2] 국립중앙도서관 오아시스 누리집 (https://nl.go.kr/oasis)

5 Metadata Object Description Schema(메타데이터 객체 기술 스키마로 디지털 자원에 특화된 표준 목록 형식)

4. 국제활동

인터넷 보존을 위한 국제기구의 탄생, IIPC

2003년 7월 프랑스 국립도서관을 중심으로 12개 국가도서관이 합심하여 국제인터넷보존컨소시엄(IIPC)을 창설하였다. IIPC는 웹 아카이빙이라는 난제를 해결하기 위해 결성된 국제기구로 국립중앙도서관은 대한민국 국가대표도서관으로 2008년부터 회원으로 가입하여 활동하고 있다. 현재 국가도서관을 중심으로 35개국 53개 기관이 참여하고 있으며, 전 세계 인터넷 정보자원의 수집 및 보존을 위한 공통의 도구·기술 등 표준 개발과 지원, 웹 콘텐츠 활용을 위한 다양한 연구, 인터넷 보존 이니셔티브 활동, 국제 정보자원 공유 등을 진행하고 있다. 국립중앙도서관은 국제 표준을 준수하고 IIPC에서 오픈소스로 제공하며 국제적으로 상용되고 있는 웹 아카이빙 도구와 기술을 오아시스에 적용하고 있다. 또한 평창동계올림픽, 코로나19 등 세계적인 이슈 관련 국제 공동 웹 아카이브 컬렉션에 참여하여 국내 관련 웹사이트 정보를 공유하고 있다.



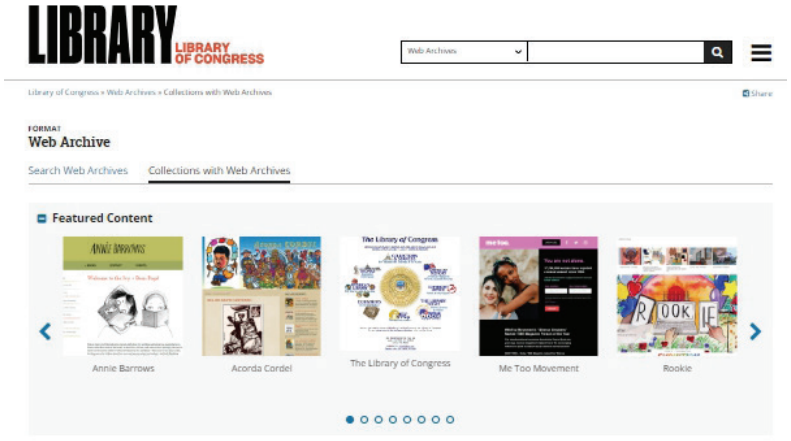
[그림 3] IIPC 웹사이트 (https://netpreserve.org)

Ⅱ. 외국 국가도서관 사례

1. 미국 의회도서관

미국 의회도서관(Library of Congress, LC)은 미네르바(MINERVA)[6] 웹 보존 프로젝트를 통해 2000년부터 웹 아카이빙을 시작하였다. 컬렉션 기반의 이벤트 또는 테마 관련 웹사이트의 선별적 수집과 정부기관 웹사이트를 대상으로 한 포괄적 수집을 병행하고 있다. LC는 전 세계 11개 국가도서관과 인터넷 아카이브(Internet Archive, IA)[7]와 함께 IIPC 초기 창립 회원으로 웹 아카이빙 공동 발전을 위한 협력을 주도적으로 진행하고 있다. 특히 지난해는 웹 아카이빙 22주년으로 웹 아카이브를 포함한 디지털 컬렉션에 대한 이용 접근성 개선 및 다양한 연구적 활용을 위해 노력하고 디지털자원 수집 확대, 대규모 웹 아카이브에 대한 접근 및 정보 제공을 위해 LC LAB[8]을 설치하여 실험적인 작업을 추진하고 있다.

LC는 연간 11억 개의 웹자원(511TB)을 수집하여 현재까지 217억 개의 웹자원을 수집·보존(2,827PB)하고 있으며, LC 웹 아카이브(Library of Congress Web Archive, LCWA)[9]를 통해 9·11 테러, 이라크전쟁, 선거 등 주제 전문가에 의해 선정된 웹 컬렉션 3만 건을 제공하고 있다. 최근에는 데이터랩을 통해 미국선거 및 미국문화 관련 대용량 웹 아카이브 데이터 세트를 제공하고 있다.



[그림 4] 미국 의회도서관 웹 아카이브 (LCWA)

6 Mapping the INternet Electronic Resources Virtual Archive. 2000년 LC에서 진행한 파일럿 웹 아카이빙 프로젝트

7 1996년 최초로 인터넷 보존을 시작, 2001년 웹페이지 재생 도구인 웨이백 머신을 개발하였으며 7,350억 개의 웹페이지 등 전 세계 온라인 디지털자원을 보존 및 제공하고 있는 비영리단체

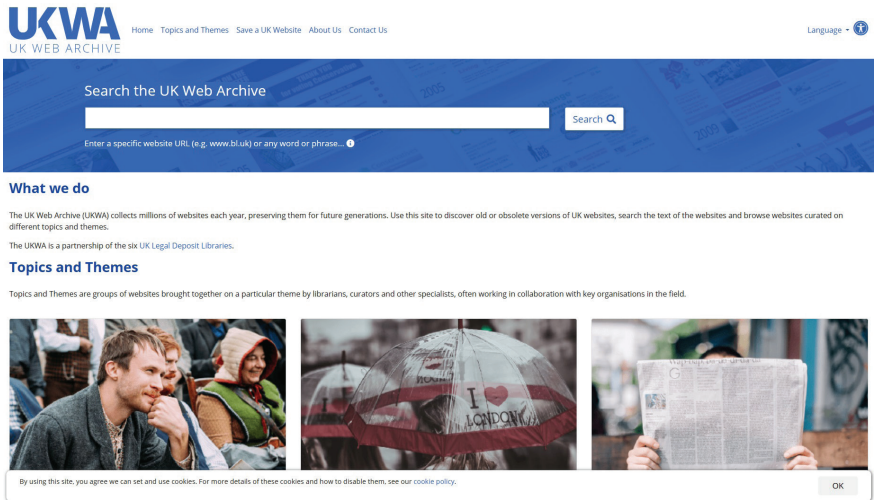
8 https://labs.loc.gov/work/experiments/webarchive-datasets/

9 https://loc.gov/web-archives/collections/

2. 영국 국립도서관

영국 국립도서관(British Library, BL)은 2005년부터 웹사이트를 수집해왔으며, 2013년부터 국가 도메인 전체를 대상으로 수집하고 있다.[10] 납본 규정에 따라 매년 영국 도메인 웹사이트 전체(수십억 개 파일을 포함한 4백만 개의 웹사이트)를 수집하고, 영국 웹 아카이브(UK Web Archive, UKWA)[11]를 통해 웹사이트 검색은 물론 영국의 생활 및 사건, 연구 관련 100개 이상의 주제 및 테마 컬렉션을 제공하고 있다. 컬렉션 대부분은 6개의 법정 납본 도서관[12]에서 이용할 수 있다. 영국 국립도서관 역시 IIPC 창립 회원으로 초국가적인 사건과 관련된 웹사이트를 보존하기 위해 전 세계적인 공동 컬렉션을 구축하는데 참여하고 있다.

특히, 1996년부터 2013년 사이에 인터넷 아카이브(IA)에서 수집한 UK 웹사이트에 대한 전문(full-text) 검색과 단어 및 문장 흐름에 대한 시각화, 파생된 데이터 세트를 제공하는 샤인(SHINE) 서비스를 제공하고 있으며, 웹 데이터 분석 서비스인 글램 워크벤치(GLAM Workbench) 서비스를 개발하여 제공하고 있다.



[그림 5] 영국 웹 아카이브 (UKWA)

10 https://bl.uk/collection-guides/uk-web-archive

11 https://webarchive.org.uk/en/ukwa/

12 British Library, National Library of Scotland, National Library of Wales, Bodleian Libraries, Cambridge University Libraries and Trinity College, Dublin

3. 프랑스 국립도서관

프랑스 국립도서관(Bibliothèque Nationale de France, BNF)은 2002년 프랑스 선거 관련 웹사이트 수집을 시작하였다. 2004년부터 인터넷 아카이브(IA)와 협업을 진행하고 2006년부터 온라인 자료 납본을 시행하면서 프랑스 도메인 웹사이트를 수집하기 시작하였다. 현재는 프랑스 도메인 웹사이트에 대한 포괄적 수집과 주제 및 이벤트 관련 웹사이트에 대한 선택적 수집을 병행하고 있다.[13]

프랑스 국립도서관은 IIPC 창립 회원으로 현재 85명 큐레이터와 프랑스 지역 도서관, 연구소, 협회, 전문기관 등 20개 이상의 파트너 기관과 협업을 통해 웹 아카이빙을 공동으로 추진하고 있다. 온라인 일기, 웹의 역사를 기록하는 활동가 웹사이트, 블로그, 문학 웹사이트 등 테마 컬렉션과 국가, 지역 및 유럽의 다양한 선거 컬렉션을 구축하였으며 연간 400만 개 사이트와 20억 개 웹페이지를 수집하여 분산보존시스템(SPAR)에 저장하고 있다. 2021년에는 디지털 컬렉션의 디지털 코퍼스 제공 및 연구 지원을 위한 데이터 세트 제공을 목표로 데이터랩을 신설하여 웹 데이터 서비스를 확대하고 있다.

[그림 6] 프랑스 국립도서관 웹 아카이브 (Archives de l'Internet)

13 https://bnf.fr/fr/archives-de-linternet

4. 일본 국립국회도서관

일본 국립국회도서관은 인터넷자료수집보존사업(Web ARchiving Project, WARP)[14]으로 2002년부터 웹사이트를 수집하기 시작하였다. 2010년 시행된 국립국회도서관법에 따라 정부, 국회, 법원, 공립대학 등 공공기관 웹사이트를 수집하고 있으며, 재단, 협회, 정당 등 사립기관과 일본에서 개최되는 문화 및 국제 행사 관련 웹사이트를 저작권자의 허가를 받아 수집하고 있다.

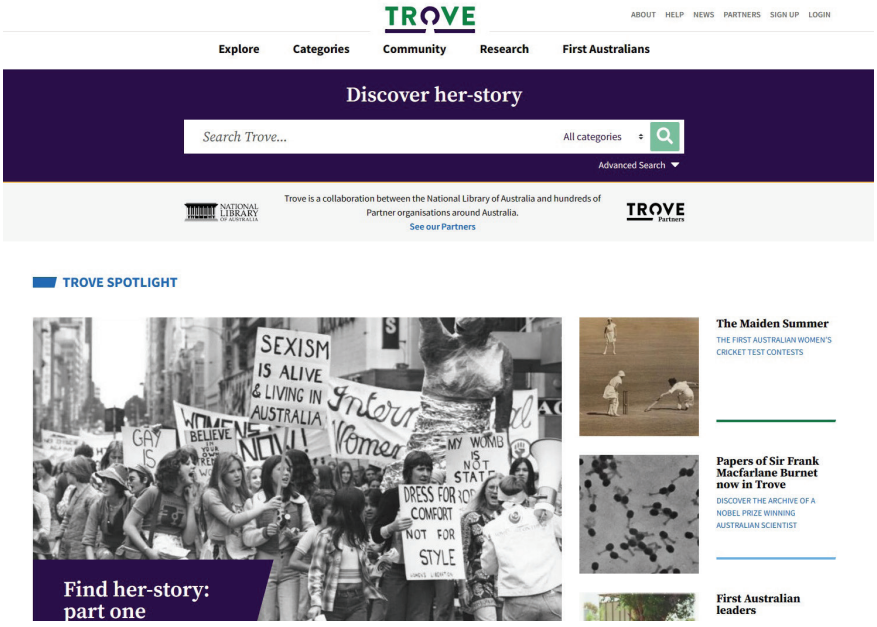
현재 122억 개 파일, 2,600TB에 달하는 웹자원을 수집 및 보존하고 있으며 데이터 마이닝에 적합한 데이터를 만드는 방법을 연구하고 웹 아카이브에 특화된 강력하고 정확한 전문검색 엔진을 개발하기 위해 노력하고 있다.

[그림 7] 일본 국립국회도서관 웹 아카이브 (WARP)

14 https://warp.da.ndl.go.jp

5. 호주 국립도서관

호주 웹 아카이브(Australia's Web Archive)는 판도라(PANDORA)[15] 프로젝트로 시작되었다. 1996년 호주 국립도서관(National Library of Australia, NLA)에서 처음으로 온라인 자료를 수집하기 시작하였고 현재 호주 전역의 9개 도서관 및 문화기관과 협력을 통해 웹사이트를 선별적으로 수집하고 있다. 주로 호주 및 호주인과 관련된 웹사이트와 웹 간행물을 수집하고 있으며, 호주의 역사와 문화, 사회, 정치 관련 웹사이트와 국가 중요 사건 관련 웹사이트를 수집하고 있다. 1960년 제정된 국립도서관법, 1968년 제정된 저작권법에 따라 매년 호주 도메인 웹사이트를 대규모로 수집하고 있다. 판도라(PANDORA)를 통해 수집된 웹 콘텐츠를 포함해서 80억 개 이상의 웹사이트를 트로브(TROVE)[16] 검색 서비스를 통해 대중에게 제공하고 있다.



[그림 8] 호주 웹 아카이브 검색 서비스 (TROVE)

15 Preserving and Accessing Networked Documentary Resources of Australia. 호주 및 호주인과 관련된 온라인 출판물을 수집하고 장기간 보존하기 위해 1996년 국립호주도서관에서 설립. http://pandora.nla.gov.au/

16 https://trove.nla.gov.au

Ⅲ. 당면 과제

1. 수집 대상 웹자원의 폭증

인터넷 등장 이후 누구나 원하는 정보를 생산, 유통, 소비할 수 있게 되었다. 실제로 가장 많은 정보가 인터넷을 통해 양산되고 있으며, 정보 채널 역시 웹사이트, 블로그, 트위터, 인스타그램, 유튜브, 메타버스(Metaverse), 가상현실(VR), 혼합현실(MR) 등 새로운 소통 수단을 통해 무한 확장되고 있다.

국가지식정보자원의 보고로서 국립중앙도서관은 기존의 인쇄자료, 비도서자료를 확충하는 전통적인 기능을 유지하면서 급증하고 있는 새로운 유형의 온라인 자료를 장서로 포함하기 위한 노력을 계속하고 있다.

여기에 수많은 웹 정보자원 중에 가치 있는 자료를 선별해야 하는 과제가 있다.[17] 국립중앙도서관은 『도서관법』 제22조(온라인 자료의 수집)에 따라 보존가치가 높은 온라인 자료를 선정하여 수집하고 있다. 수집 기준에 따라 공공 웹자원을 우선 수집하지만, 실제 개인용 또는 상업용 웹자원이 대부분을 차지하므로 보존 가치를 판단해야 하는 어려움이 있다. 또한 웹자원은 재생산이 가능하므로 원본과 편집본을 검증해야 하며 진본성을 보장해야 하는 문제가 있다.

웹은 방대하고 복잡하며 계속 갱신되는 동적인 자원으로 수집하기가 쉽지 않다. 현재 웹 크롤러를 개발하여 주기적으로 수집하고 있지만 모든 웹자원을 수집하기 어려울 뿐만 아니라 완전하게 수집하는 데에도 어려움이 있다. 현재 웹사이트 기준으로 70% 이상 수집이 되면 수집에 성공한 것으로 판단하고 있으며 최대 5깊이(depth)까지 수집을 제한하고 있다. 웹사이트 수집은 완전하게 구현하기 어려워 무결성을 보장하기 어렵고 품질에 한계가 있다. 외국의 경우 사이트별로 최적화하는 작업을 진행하기도 하는데 이를 위해서는 제반 여건이 마련되어야 한다.

2. 시스템 과부하 및 보안의 문제

웹자원은 대용량 자원으로 수집, 보존, 서비스까지 관리를 위한 시스템 지원이 필수적이다. 국립중앙도서관은 오픈 소스 기반의 시스템을 구축하여 사용하고 있다. 현재 245만 건(1,007TB)에 달하는 대용량 웹자원이 축적되어 있다. 수집 서버, 스토리지, 네트워크 등 상당한 자원을 할당하여 운영하고

17 국립중앙도서관 온라인자료과 (2022). 미래 도서관 웹 콘텐츠 수집 정책 : 제59회 전국도서관대회 세미나. 서울: 국립중앙도서관 온라인자료과.

있는데, 이로 인한 시스템 과부하로 원활하게 관리하기 어려운 상황이다. 실제 수집에 상당한 시간이 소요되고 지연이 발생하고 있어 검증하는 과정은 물론 오아시스 누리집을 통한 검색 및 웨이백(wayback)[18] 재생까지 신속한 처리가 어렵고 다소 불안정한 상태로 서비스되고 있다. 그리고 웹사이트 전체를 깊이 있게 수집하므로 악성코드가 유입될 가능성이 있고 개인정보가 포함될 수 있어 보안에 취약한 약점이 있다.

3. 예산 및 조직의 한계

국립중앙도서관의 오아시스 사업은 국가 웹자원을 수집 및 보존하는 선도적인 사업으로 사업 초기에는 온라인 자료 수집 및 메타데이터 DB 구축 사업에 포함되어 사업 기반 조성 및 시스템 구축을 위해 상당한 예산이 지원되었다. 그러나 2016년부터 웹사이트 아카이빙 사업이 별도로 분리되면서 사업 예산이 급감하였다. 현재 1.3억 원이라는 한정된 예산으로 국가 웹자원을 수집 및 보존, 서비스해야 하는 상황으로 사업 유지조차 어려운 실정이다. 시스템 유지·운영을 위한 별도 예산이 있지만 비중이 크지 않다.

웹자원은 가장 복합적인 형태를 지닌 정보자원이며 새로운 유형의 웹자원이 출현하고 진화하고 있어 수집·보존·서비스까지 끊임없는 연구 개발 및 기술 지원이 필요하다. 또한 국가 도메인 전체와 최근 성장하고 있는 글로벌 K-웹자원까지 국가 웹자원의 수집 범위가 지속적으로 확대되고 있다. 하지만 제한된 예산과 조직 등 여건 부족으로 인하여 소멸되기 전의 웹자원을 단순 수집하는 데 치중하고 있다.

Ⅳ. 오아시스 발전을 위한 제언

1. 인식 개선 및 활용도 제고

국가도서관은 모든 정보자원을 수집 및 보존, 제공해야 하는 사명이 있으며, 현시대의 가장 강력한 정보자원인 웹자원 역시 수집 및 보존하고 제공해야 하는 당위성을 지니고 있다. 국립중앙도서관은 국가대표도서관으로 전 세계 도서관과 함께 웹 아카이빙이라는 도전적인 과업을 실현하고 있다. IIPC 회원으로 참여하며 다양한 정보와 기술을 습득하고 선진 사례를 벤치마킹하고 있다. 따라서 이제는 웹

18 WARC 파일 포맷으로 저장된 웹 페이지들을 보기 위해 인터넷 아카이브(IA)가 개발한 디지털 타임 캡슐

아카이빙의 필요성을 적극적으로 피력해야 한다.

완전한 수집에 한계가 있지만 전 세계 국가도서관을 중심으로 쉽게 생성·소멸되는 웹자원을 제대로 수집 및 보존, 활용하기 위해 다양한 창의적인 실험이 진행되고 있다. 국립중앙도서관은 웹자원을 유용한 자원으로 재가공하여 제공해야 하며 웹 데이터의 활용 가치를 보여주어야 한다. 실제로 미국, 프랑스 등 선진 도서관은 완벽하게 재생되는 웹 아카이브 컬렉션을 제공하고 있으며 최근에는 국가도서관 연구소(LAB)를 통해 대용량의 웹 데이터를 가공해서 제공하고 있다.

그리고 웹자원에 대한 개방성 확대와 아카이브 활용 방안을 다각적으로 고민해야 한다. 오아시스는 사라지는 과거와 현재의 웹자원을 유일하게 보존하여 후대에 전승하는 미래형 타임캡슐로써 의의가 있다. 오아시스의 중요성에 대한 홍보를 강화하고 오아시스 자원의 대국민 서비스 제공을 위한 저작권자의 동의를 적극적으로 독려하여 외부 공개 자원의 범위를 확대해야 한다. 또한 실생활에 도움이 되는 유용한 웹자원 분석 및 시각화 서비스를 개발해서 제공해야 한다. 아울러 방대한 웹 데이터에 대한 장벽 없는 개방을 통해 각계각층의 연구에 활용될 수 있도록 지원해야 한다. 이를 통해 웹 아카이브에 대한 인식 개선은 물론 무궁무진한 웹자원의 활용 가치를 증명할 수 있을 것이다.

2. 미래 자원 수집 및 보존을 위한 협업과 지원 확대

도서관의 미래 자원으로 웹 아카이빙이라는 고난도 과업을 수행하기 위해서는 실질적인 관심과 지원이 필요하다. 실제로 웹자원의 유형이 다양화되고 있으며 수집해야 하는 대상도 확대되고 있다.

국립중앙도서관은 주요 기관 및 중요 주제 관련 웹자원에 대한 선별적 수집과 국가 도메인 웹사이트에 대한 포괄적 수집을 병행하고 있으며, 지난해 부터 한류 확산과 더불어 중요성이 부각되고 있는 글로벌 K-웹사이트를 시범적으로 수집하는 등 웹자원의 수집 범위를 확대하고 있다. 또한 블로그, 유튜브, 누리소통망 서비스(SNS), 메타버스, 가상현실, 혼합현실 등 새로운 유형의 웹자원을 수집 및 보존하기 위해 시도하고 있다.

그러나 오아시스 시스템은 오픈 소스 기반의 시스템으로 구축되어 처리속도와 성능에 한계가 있고, 오아시스 업무 역시 최소한의 예산과 인력으로 운영되고 있어 제반 여건이 부족한 상황이다. 방대하고 복잡한 웹자원을 원활히 수집·보존·서비스하기 위해서는 예산 및 인력, 시스템(H/W, S/W) 등 전반적인 인프라 지원이 확대되어야 한다. 더불어 수집부터 서비스까지 연구 개발과 기술 지원이 필요하며 사서,

정보기술자, 전문가 그룹 등으로 구성된 팀 단위 조직과 장단기적인 협업이 필요하다. IIPC에서는 웹 표준에 따라 웹자원을 안정적으로 수집 및 보존하기 위한 다양한 도구와 소프트웨어 개발하고 있으며 오픈 소스로 제공하고 있다. 따라서 이러한 도구와 소프트웨어를 도입할 수 있는 여건을 조성하고 오아시스 사업과 시스템에 대한 고도화를 고려해야 한다.

3. 국가 웹 아카이브 서비스 구현

국립중앙도서관은 2006년부터 오아시스 누리집을 구축하여 대국민 웹자원 검색 및 이용 서비스를 제공하고 있다. 특히 지금은 사라진 과거의 웹사이트를 오아시스에서 찾아볼 수 있으며 웹사이트의 변경 기록(메멘토)과 변천사를 시간순으로 확인할 수 있다. 삼풍백화점 붕괴 사고 이후 최근 이태원 참사까지 관련 웹자원을 『국가재난아카이브』에서 확인할 수 있으며 선거, 올림픽, 월드컵 등 국가 중요 행사 관련 웹자원을 『주제·이슈 컬렉션』에서 확인할 수 있다. 누리 소통망 서비스(SNS)의 경우 정부 및 공공기관 트위터를 수집해서 제공하고 있으며, 웹 데이터 분석 및 시각화 서비스인 태그 클라우드, 웹 트렌드 서비스를 개발하여 제공하고 있다.

국립중앙도서관은 국내에서 유일하게 웹자원 아카이브(OASIS)를 구축하여 제공하고 있지만 아직 미진한 점이 많다. 장기적으로 다양한 유형의 웹자원을 체계적으로 수집·보존·활용할 수 있는 국가 단위의 플랫폼으로 개선되어야 한다. 네이버, 구글 등 상용 엔진과 같은 고성능 검색엔진을 적용하여 과거의 웹자원을 편리하게 검색하고 이용할 수 있어야 한다. 인공지능 기반 웹 데이터 분석 및 활용 시스템을 도입해서 방대한 웹에서 원하는 정보를 신속하고 정확하게 찾아주는 지능형 서비스가 필요하고 주제별·시대별 웹 데이터의 흐름을 보여주는 고도화된 웹 데이터 분석 및 시각화 서비스가 구현되어야 한다.

원자료인 국가 웹자원의 활용 가치는 무궁무진하다. 우리나라 웹자원 전체를 모아놓고 체계적으로 정리해서 제공하는 국가 웹 아카이브로 구현되어야 한다. 미래 세대를 위해 국가도서관만이 해낼 수 있는 중요한 과제이다.

참고문헌

국립중앙도서관 온라인자료과 (2022). 미래 도서관 웹 콘텐츠 수집 정책 : 제59회 전국도서관대회 세미나. 서울: 국립중앙도서관 온라인자료과.

유네스코한국위원회 (2003). 디지털 유산의 보존에 관한 헌장. 출처: https://unescokor.cafe24.com/assets/data/standard/KsGRkyu8gJyKSiQWjoMcarvOshm6Ly_1217257200_1.pdf