웹 기록물 아카이빙을 위한 워크플로우 및 메타데이터 연구

차승준, 천동석, 이규철†
충남대학교 컴퓨터공학과
e-mail:(junii, ikarus1004, kclee)@cnu.ac.kr

A Study of the Workflow and the Metadata for Web Records Archiving

Seung-Jun Cha, Dong-Suk Chun, Kyu-Chul Lee
Dept of Computer Science, Chungnam National University

요 약

웹은 급속하게 변화하는 현대사회에서 정부와 시민들의 주요 의사소통의 채널이 되고 있다. 웹에서 유통되는 정보량이 급증하면서 정보원으로서의 웹에 대한 의존도가 크게 높아졌을 뿐만 아니라 전적으로 웹에만 존재하는 정보자원도 증가하고 있다. 중요한 가치를 지닌 웹사이트는 짧은 수명주기와 수집, 보존, 활용에 대한 방안이 없어 소멸되고 있는 실정이다. 이러한 문제를 해결하기 위해 웹 기록물 아카이빙을 위한 기반기술로 워크플로우 및 메타데이터 정의가 필요하다. 따라서 본 논문에서는 웹 기록물을 아카이빙하기 위해 선별, 수집, 품질관리 및 목록화, 보존, 저장으로 구성되는 워크플로우 및 장기 보존과 검색에 필수적인 메타데이터를 정의하였다. 이러한 연구 개발 및 적용을 통해 사라져 가는 중요한 자원인 웹 기록물을 후대에 중요한 기록물 자원으로 저장 및 관리할 수 있게 될 것이다.

1. 서론

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, 전자저널과 같은 형태로 연구 결과물을 배포하고 획득한다. 웹이 출현한지 얼마 지나지 않아서 웹은 연구기관의 범위를 벗어나서 정부, 기업과 상업, 교육, 언론과 출판, 그리고 개인 영역으로 확대되었다.

웹사이트에서 유통되는 정보량이 급증하면서 정보원으로서의 웹에 대한 의존도가 크게 높아졌을 뿐만 아니라 전적으로 웹에만 존재하는 정보자원도 증가하고 있다.[1] 그 가운데 잘못된 정보를 담고 있거나 아주 일시적인 가치만 가진 것들도 있는 반면에 역사적·문화적·학술적 가치와 법정 증거능력을 가지고 있기 때문에 장기간 보존해야 할 것들도 많이 있다.

웹사이트는 그 자체로서 지식·정보의 저장소이며 디지털 문화의 자산으로서 중요한 가치를 지니고 있음에도 불구하고, 웹이라는 짧은 생명주기 및 수집, 보존, 활용에 대한 방안이 없어 소멸되고 있는 실정이다. 하루에도 수많은 웹 기록물이 생성되었다가 소멸되어지고 있다

이러한 문제를 해결하기 위해서 웹 기록물 아카이빙에 대한 기반 기술의 연구가 필요하다. 본 논문에서는 웹 기록물 아카이빙을 위한 워크플로우에 대해 정의하고, 장기보존 및 목록화를 위한 메타데이터 정보를 설명한다. 이를 통해 웹 기록물도 중요한 자원으로 후대까지 보존 및 제공할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 아카이빙을 수행하는 다른 프로젝트에 대해 설명하고 3장에서는 웹 기록물 아카이빙에 대한 워크플로우를 정의한다. 4장에서는 저장되어야 하는 메타데이터에 대해 설명하며 5장은 결론 및 향후 연구에 대해 설명한다.

2. 관련연구

1994년 캐나다 국립도서관(National Library of Canada)의 EPPP(Electronic Publications Pilot Project)가 시작된 이후 웹 자원에 대한 관심이 여러 국가도서관으로 확산되었다. 특히 WAIS를 개발한 Brewster Kahle이 동료와 함께 1996년 4월에 인터넷 아카이브(Internet Archive)를 설립한 것은 웹 자원의 아카이빙에 관한 대중적인 관심을 이끌어 냄으로서 각 국의 프로젝트 추진에 큰 힘이 되었

† 교신저자

<표 1> 웹 기록물 아카이빙 프로젝트 현황

국가	사업명	추진주체	수집방법	접근성	규모
호주	PANDORA	호주 국립도서관	선택	공개	353Gb
영국	Britain on the Web(Domain UK)	영국 국립도서관	선택	비공개	30Mb
일본	WARP	일본 국립국회도서관	선택	공개	524Gb
미국	MINERVA	미국 의회도서관	선택	비공개	35사이트
덴마크	netarchive.dk	Royal Library와 The state and University Library	선택	비공개	280Gb
프랑스	BnF Web Archiving initiative	Biblothque nationale de France	선택/포괄	비공개	1 Tb
노르웨이	PARADIGMA	노르웨이 국립도서관	선택/포괄	제한적 공개	140Gb
스웨덴	Kulturarw3	Koninklijke bibloteket(KB)	포괄	제한적 공개	6Tb
핀란드	EVA	핀란드 국립도서관	포괄	비공개	401Gb
오스트리아	AOLA	ONB/TY Wien	포괄	비공개	448Gb
미국	Internet Archive	Internet Archive	포괄	공개	150Tb

다. 인터넷 아카이브는 처음부터 공공자료를 수집하여 보존하고 역사가, 연구자, 학자 등에게 장기적으로 이용시키는 디지털도서관을 표방하였다.[2)]

이후로 지금까지 웹 자원의 수집과 보존을 위한 시도가 다양한 형태로 이루어지고 있다. <표 1>은 그 중 대표적인 사례를 요약한 것이다.

이하에는 대표적인 사례로 PANDORA, MINERVA, IIPC에 대해서 간략히 소개하기로 한다.

2.1 PANDORA

PANDORA는 호주 국립도서관(National Library of Australia)에서 시작하였다.[5)] 온라인 출판물은 사회적, 지적, 문화적자원으로 중요하다는 관점에서 1996년 프로젝트로 설립되었다. 2000년 시드니 올림픽 관련자료 수집에 이용되기도 했다. 2008년 6월에 아카빙된 자료는 <표 2>과 같다

<표 2> PANDORA 아카이빙 현황

	이번달	지난달	증감
아카이빙된 제목의 수	19,257	18,874	383
아카이빙된 인스턴스 수	38,060	37,422	638
파일의 수	51,469,417	40,518,371	951,046
데이터의 크기	2.19 TB		

2.2 MINERVA

MINERVA(Mapping the Internet Electronic Resources Virtual Archive)는 미국 의회 도서관(Library of Congress)에서 주도 하였다.[6)] MINERVA에는 대통령선거, 911 테러, 이라크전쟁 등 특정 주제를 다루며 일정한 선택기준을 만족시키며, 자유롭게 접근할 수 있는 사이트를 도서관 직원이 선택하여 수집한다. 또한 사이트 권리 보유자와 특약협정을 맺고 공개조건을 개별적으로 설정한다.

2.3 IIPC

IIPC(International Internet Preservation Consortium)는 2003년 설립된 12 회원사를 가지고 있는 국제 웹 아카이빙 단체로 다음과 같은 목표를 가지고 있다.[7)]

- 장기보존될 수 있도록 인터넷 컨텐츠의 자원들을 수집 가능하도록 한다.
- 국제적인 장기보존이 가능하도록 개발과 일반적인 도구, 기술과 표준의 사용을 장려한다.
- 인터넷 아카이빙과 보존을 하고자 하는 국립 도서관들을 장려하고 지원한다.

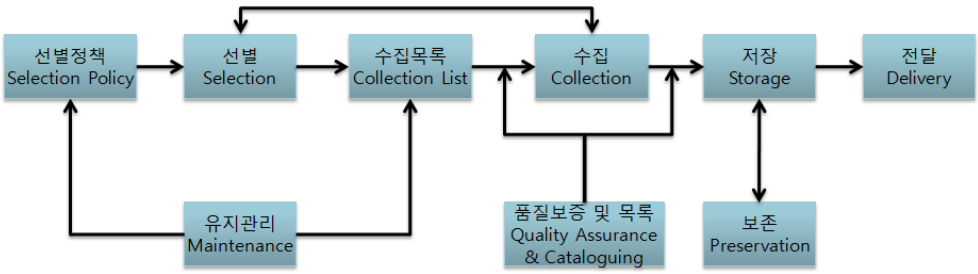
3. 워크플로우 정의

웹 기록물 아카이빙을 위한 워크플로우는 (그림 1)과 같다. 적절하게 정해진 선별 정책에 의해 선별된 사이트를 정해진 수집방법에 의해 수집된다. 이렇게 수집된 기록물은 품질보증 및 목록에 의해 진본성을 유지하며 장기간의 보존을 제공하며 사용자에게 전달해준다.

3.1 선별

웹 기록물 아카이빙에 있어 적당한 선별정책을 결정하는 것은 반드시 선행되어야 할 사항이다. 선별정책의 결정 요소에는 수집 기관의 목표, 지적 재산권 관련, 기구 자원의 존재성 있다. 현재 웹 기록물 아카이빙은 도서관, 박물관, 연구기관, 상업기구 등 다양한 곳에서 시행되고 있으며 각각마다 알맞은 선별 방법을 사용해야 한다.

선택의 범위에 따라 다른 접근법들이 존재한다. 크게 비선별적(Unselective), 주제적(Thematic), 선별적(Selective)으로 나눌 수 있다.



(그림 1) 웹 기록물 아카이빙 워크플로우

3.2 수집방법 및 수집

수집방법으로는 크게 컨텐츠-기반 수집(Content-driven collection)과 이벤트-기반 수집(Event-driven collection)으로 나눌 수 있다. 컨텐츠-기반 수집은 컨텐츠에 기반하여, 즉 웹 사이트의 기본적인 컨텐츠를 아카이빙하기 위한 방법이다. 이벤트-기반 수집(Event-driven collection)은 웹 서버와 브라우저 사이에 발생된 실제적인 트랜잭션(transaction)을 처리하는 것이다. 각 수집방법에 대한 자세한 내용은 <표 3>와 같다.

<표 3> 수집방법

	컨텐츠-기반	이벤트-기반
클라이언트	원격 하베스팅	
서버	직접전송 데이터베이스 아카이빙	처리행위 아카이빙

직접전송(Direct Transfer)는 가장 간단한 방법으로 웹 서버에서 직접 그 원본 데이타를 복사하여 가져오는 방식이다. 이를 위해서는 웹 사이트의 관리자와의 협력이 필요하다.

원격 하베스팅(Remote Havesting)은 가장 널리 사용되는 방법으로 웹 크롤러 소프트웨어를 사용하여 웹 서버로부터 컨텐츠를 획득하는 방식이다.

데이터베이스 아카이빙(Database Archiving)은 데이터베이스기반 사이트를 아카이빙 하는 것으로 다음 단계로 이루어진다. 먼저 아카이브된 데이터베이스를 위한 저장소를 표준 데이터 모델과 형식으로 규정하고 각 데이터베이스를 표준 형식으로 전환한다. 이런 방식으로 아카이브된 데이터베이스에 표준 접근 인터페이스를 제공한다.

처리행위 아카이빙(Transaction Archiving)은 웹서버로부터 전달되는 콘텐츠보다는 웹 서버와 브라우저 사이에 발생하는 실제 처리행위(transactions)를 수집하는데 초점을 맞추는 것이다.

3.3 품질보증 및 목록

품질보증(Quality assurance)은 웹 기록물 아카이빙 절차에 필수적인 구성 요소이다. 품질보증의 본질 및 등급은 요구사항과 수집하기 원하는 리소스에 많이 좌우된다

목록(Cataloguing)은 전체 웹 아카이빙 단계에서 어떠한 단계에서 발생되는가와 그 중요성에 대해 설명한다.

3.4 보존

보존의 목적은 대상의 가치가 유지되면서, 영속적인 접근을 보장하는 것이다. 즉, 성공적인 보존이란 사용자들의 접근이 가능하며 본래의 가치를 그래도 보존하여 전달해야 하는 것이다. 이러한 웹 기록물의 보존은 디지털 대상의 보존과 유사하여 같은 기술이 적용된다. 하지만 이러한 보존은 아카이빙이 시작된 1996년에 이루어졌지만 여전히 초기단계에 머물러 있다.

3.5 전달

아카이빙된 웹 기록물에 대한 발견이나 확인에 있어 기본적인 메소드는 크게 검색 접근(Searching access)과 열람 접근(Browsing access)이 있다.

검색 접근은 사용자에게 필수적인 툴을 제공하는 것으로 작게는 제목이나 URL의 검색부터 크게는 전체 콜렉션에 걸친 전문 검색(full-text search)을 제공한다.

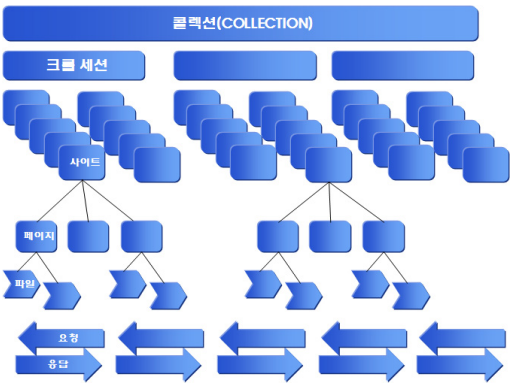
열람 접근(Browse access)은 미리 정해진 계층구조로 자료들이 분류되어 있어 사용자들은 “drill-down”식으로 접근 가능하다. 정해진 분류체계가 사용자의 예상과 같으면 매우 유용하지만, 사용자의 그룹이 많기 때문에 어려운 점이 있다.

4. 메타데이터

웹 기록물에 대한 설명정보나 메타데이터의 보존은 디지털 자원의 관리에 있어 기본적인 필요사항이다. 특히 장기의 보존을 위해 추가적인 기술적 메타데이터가 필요하다. 이는 수동적, 능동적 보존 절차를 모두 지원해야 해야 한다.

이를위해 아카이빙된 컨텐츠와 컨텐츠에 대한 설명정보로 구성된 메타데이터를 함께 저장해야 한다. 웹 기록물에 대한 메타데이터는 설명정보의 필요성이 덜 중요하다. 이

는 웹 기록물 아카이빙은 자동적 수집이 많기 때문에 일일이 많은 설명정보를 기술하기 어려움이 있기 때문이다. 따라서 더블린 코어(Dublin Core) 수준의 설명정보를 공통적으로 사용된다.



(그림 2) 메타데이터 적용 수준

기술적 메타데이터 적용 수준은 (그림 2)와 같다. 아카이빙된 콜렉션(Collection)안에는 각각 크롤링된 세션들이 존재한다. 각 세션안에는 여러 가지 사이트들이 저장되어 있다. 각 사이트에는 여러 페이지와 페이지를 구성하고 있는 파일들이 있다. 또한 사이트에 대한 정보를 보기 위한 요청과 응답메시지들도 함께 저장되어 있다. 이러한 적용 수준에 맞춰 다음과 같은 기술적 메타데이터가 저장되어야 한다.

4.1 문서와 관련된 메타데이터

문서와 관련된 메타데이터로는 원본 위치(Original location), 크기(Size), 수집일(Ingest Date)이 있다. 원본 위치는 아카이빙을 수행한 위치로 URI형식으로 기술된다. 크기는 아카이빙된 총 사이즈를 의미하고, 수집일은 아카이빙이 수행된 날짜를 의미한다.

4.2 트랜잭션 행위

트랜잭션 행위는 해당 크롤링에 대한 정보로 크롤러 정보 (type, set-up, IP)와 서버 정보(IP, DNS information)로 구성된다. 크롤러 정보로 크롤링 세션에 대한 정보, 크롤링 머신 IP주소, 서버IP 주소가 저장되며 서버 정보로 서버 IP 주소와 도메인 이름 정보(whois 정보)가 같이 저장되어야 한다.

4.3 선별절차 관련정보

선별절차에 관련정보로 시작점 리스트(Entry Point List), 문서까지의 경로(Path to document), 시작점으로부터의 홉(Hop from entry point), 선별정책 문서(Selection policy documentation), 특정/측정(Characterization/Evaluation)에 대한 값과 도구이다.

5. 결론

본 논문은 중요성을 가지고 있지만 사라져가는 웹 기록물을 보존하기 위한 웹 기록물 아카이빙에 대한 워크플로우 정의 및 저장시 필요한 메타데이터에 대해 기술했다.

워크플로우와 메타데이터를 정의하기 위해 우선 웹 기록물 아카이빙을 이미 수행하고 있는 다른 프로젝트들에 대해 비교/분석하였다. 대표적으로 IIPC, 호주의 PANDORA, 미국의 MINERVA가 있다.

이러한 프로젝트를 분석하여 워크플로우를 정의하였다. 선별정책을 통한 선별, 선별된 사이트의 수집, 품질보증 및 목록, 보존, 그리고 전달의 단계를 거쳐서 웹 기록물 아카이빙을 수행한다.

장기적인 보존을 위한 메타데이터에 대해 설명하고 추가적으로 기술해야하는 메타데이터를 정의하였다. 웹 기록물의 특성상 기술적인 메타데이터 보다는 설명적인 메타데이터가 더 중요하다. 따라서 설명적인 메타데이터인 더블린 코어를 기반으로 추가적인 기술적인 요소를 정의하였다.

이러한 워크플로우 및 메타데이터를 통해 가치를 가지고 있는 웹 기록물들에 대한 장기적인 보존 및 필요한 사용자에게 전달을 수행할 수 있다.

하지만,실제로 이러한 내용을 적용하기 위한 추가적인 연구가 필요하다. 아카이빙의 목적에 따른 세부적인 선별정책의 적용이 필요할 것이며 또한 수집을 위한 수집기의 개발이 필요하다. 이러한 수집기는 정의한 메타데이터의 요소를 모두 포함하는 것이어야 한다.

6. Acknowledge

본 연구는 지식경제부 및 정보통신연구진흥원의대학 IT 연구센터 육성·지원사업 (IITA-2008-C1090-0801-0031)의 연구결과로 수행되었습니다. 또한 본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이에 감사드린다.

참고문헌

[1] 서혜란, “웹 아카이빙의 성과와 과제”, 2004, 한국비블리아 학회

[2] 이재윤, “국외 디지털 아카이빙 사례”, 한국교육학술정보원

[3] Adrian Brown, "Archiving Website", facet pulishing

[4] 이치주, “온라인 연속간행자료 수집 및 보존에 관한 연구”, 연세대학교 문헌정보학과

[5] IIPC, "http://www.netpreserve.org/about/index.php"

[6] MINERVA "http://www.loc.gov/minerva/"

[7] PANDORA "http://pandora.nla.gov.au/"