논문

원격수집 방식의 웹기록물 관리를 위한 웹수집기 성능 비교 연구*

Comparison of Web Crawler Performance for Web Record Management

장진호(Chang Jinho)** · 권혁상(Kwon Hyuksang)***
이규모(Lee Kyumo)**** · 최동준(Choi Dong Joon)*****

* 본 논문은 국가기록원 2022년 국가기록관리 활용기술 연구개발 사업에 의하여 연구되었음.
** 이씨플라자 지식서비스본부 본부장(prodigy@ecplaza.net)(제1저자).
*** 이씨플라자 지식서비스본부 차장(hyuksang78@ecplaza.net)(제2저자).
**** 이씨플라자 지식서비스본부 차장(scalelee@ecplaza.net)(제3저자).
***** 한국무역정보통신 전자문서실 실장, 박사(logisbrain@naver.com)(교신저자).
■투고일: 2022년 09월 30일 ■최초심사일: 2022년 10월 04일 ■최종확정일: 2022년 10월 22일.
■기록학연구 74, 155-186, 2022, https://doi.org/10.20923/kjas.2022.74.155

〈초록〉

2022년 현재 행정안전부 정부24 웹사이트에 등록된 공공부문 인터넷 웹사이트는 1만 7천여 개이다. 이처럼 많은 웹사이트를 기록으로 관리하기 위해 기록물 생산기관과 기록물관리기관이 상호 간 직접 이관하는 방식은 많은 인적·물적자원을 필요로 한다. 각 웹사이트 구동에 필요한 운영 소프트웨어와 응용소프트웨어 기술을 기록물관리기관이 보유하고 운영하는 것도 현실적으로 어렵다. 이러한 현실적 한계를 극복하기 위해 웹수집 소프트웨어를 이용하여 원격지에서 웹사이트를 자동으로 수집하는 방식이 국내외에서 사용되고 있다.

이 연구는 공공부문 인터넷 웹사이트를 원격으로 수집하여 기록으로 관리할 때 필요한 웹수집기의 성능을 비교하였다. 선행연구 및 문헌조사 등에서 검토한 다수의 웹수집 소프트웨어에 대하여 단계별 검토를 거쳐 가장 적합한 웹수집기를 선정하였다. 성능 평가 과정에는 일부 공공기관 웹사이트를 대상으로 실제 원격 수집 성능을 비교하였다. 이 연구 결과는 웹기록 관리를 위해 웹수집기 선택이 필요한 기관에 실증적이고 구체적인 성능 비교 정보를 제공한다.

주제어 : 웹기록물, 원격 수집, 웹수집기 성능 비교, Heritrix

〈Abstract〉

As of 2022, the number of Internet sites for public institutions registered on the 'Government 24' website (www.gov.kr) of the Ministry of the Interior and Safety is 17,000. The direct transfer takes a lot of human and material resources and time between the records-producing institution and the records-management institution that manages websites as records. In addition, it is practically difficult for records management institutions to migrate and operate various software and application technologies required to run each website. A method of automatically collecting websites from a remote location using web crawler software is

used domestically and abroad to overcome these practical limitations. This study compared the performance of the web crawler required to collect and manage public Internet websites as records remotely. The most suitable web crawler was selected through a step-by-step review of several web crawlers from previous studies and other literature. Several public agency websites were applied to compare the actual performance of the crawlers in the evaluation process. The study provides empirical and specific performance comparison information for organizations that need to choose a web crawler.

Keywords : Web Record Management, Remote Collection Method, Web Crawler Performance Comparison

1. 서론

1) 연구의 배경 및 목적

인터넷 웹사이트는 이 시대 가장 보편적인 정보전달 수단 중 하나이다. 공공부문에서는 중요 인터넷 웹사이트를 기록으로 관리하고 있다. 이를 웹기록물 관리라고 한다. 웹기록은 현재의 등록시스템으로는 통제하기 어려운 기록이기에 다른 방식으로 아카이빙을 한다(이영남, 2018)[1]. 공공부문의 웹기록물 관리는 2008년 정부 조직 개편 시 진행되었는데, 당시 폐지·한시 기관의 웹사이트를 국가기록원이 이관받았다. 이후 국가기록원은 2011년에서 2012년까지 생산기관 웹사이트를 원격수집하여 웹기록물로 저장, 관리하는 웹기록물 관리시스템을 구축하고 2014년부터 본격적으로 중앙행정기관 등 인터넷 웹사이트를 수집했다.

1) 이영남. (2018). 국가기록혁신과 기록담론, 기록학연구, (56), p.75

최근 공공부문 웹사이트는 서버 측에서 동작하던 웹페이지 전환이 사용자 영역에서 구동되는 스크립트 방식으로 전환되고 웹페이지가 동적으로 로딩되는 외부 라이브러리 사용이 증가하는 변화가 발생했다. 울트라 와이드 모니터, 다양한 크기의 모바일 기기 등 사용자 환경의 하드웨어가 다양화됨에 따라 화면 크기에 자동으로 맞추어지는 반응형 웹, 손가락 터치로 화면전환을 지원하는 스크립트 등 새로운 웹사이트 구축 기술이 적용되면서 정상적으로 수집되지 않는 웹사이트가 증가하고 있다.

이 연구는 기술 발전 등 변화하는 인터넷 환경에서 웹사이트를 원격 수집하는 소프트웨어인 웹수집기의 성능을 비교하고 실제 공공기관 웹사이트를 대상으로 실증 테스트 결과를 제시함으로써, 웹사이트를 기록으로 관리하고자 하는 기관에 필요한 정보 제공을 목적으로 한다.

2) 연구의 범위 및 방법

이 연구는 부문별로 4개 영역으로 구분되어 있다. 첫 번째는 웹사이트의 기록화에 관하여 선행연구를 고찰하여 웹기록물 수집 방법, 수집 도구, 보존 포맷에 대하여 검토하였다. 두 번째는 연구설계로 웹수집기 선정 방법과 비교 기준을 수립하였다. 세 번째는 다양한 웹수집기 중 적합도를 평가하여 테스트할 웹수집기를 선정하고, 웹사이트 수집 실증 테스트를 진행하였다. 마지막으로 이 연구의 결론을 요약하고 연구 한계 순서로 진행하였다.

〈표 1〉 연구 방법

(2장) 웹사이트 기록화 방안	→	(3장)연구설계	→	(4장) 웹수집기 선정 및 성능 비교	→	(5장)결론
· 웹사이트의 개념과 웹기록물의 범주 · 선행연구 고찰		·선정 방법 ·비교 기준		·라운드별 적합도 평가 ·웹사이트 수집 실증		·결론 ·연구한계

2. 웹사이트 기록화 방안

1) 웹사이트의 개념과 웹기록물의 범주

법률상 ‘웹사이트’라는 용어는 「공공기록물관리에관한법률 시행령」 제2조(정의) 제10호에 ‘웹기록물’ 정의와 「도서관법 시행령」 제13조의2(온라인 자료의 수집) 제1항에서 찾을 수 있다. 공공기록물법은 웹기록물을 ‘웹을 기반으로 생산된 기록정보자료와 웹사이트 운영 및 구축 관련 관리정보’로 정의하는 것과 달리 「도서관법」의 ‘온라인 자료’는 ‘웹사이트와 웹자료 등’으로 구분하고 있어 운영 관련 정보 및 산출물 등을 포함한 「공공기록물법」의 ‘웹기록물’이 도서관법의 ‘온라인 자료’보다 더 큰 범위를 규정하고 있다. 웹사이트는 웹브라우저 기반으로 접속하여 이용하는 방식이 일반적이나 전용 접속 프로그램이나 전용 앱 등 웹브라우저 이외의 도구를 이용하기도 한다. 이지은(2006)의 연구에서는 오스트레일리아 국립문서국(NAA, National Archives of Australia)의 분류를 인용하여 웹사이트를 인트라넷, 익스트라넷, 공공웹사이트 3가지로 재분류하고 용도를 설명했다. Wanda Archy(2018)의 온라인 칼럼에서는 웹사이트를 표면웹, 심층웹, 다크웹으로 구분했다. 표면웹은 검색엔진 등에서 클릭하여 접근할 수 있는 웹사이트 및 소셜미디어 등을 포함하며, 심층웹은 검색엔진 등으로 접근은 가능하나, 정보의 양이 방대하여 비용 등을 지불하고 이용해야 하는 구독 웹사이트로 구분했다. 마지막으로 다크웹은 기존 검색엔진으로 접속할 수 없는 웹의 일부로 익명화된 소프트웨어가 있어야 접속 가능하다고 설명했다. 위 연구에서 웹사이트는 내부용과 외부용으로 구분되거나, 데이터 중심의 웹사이트 등 서비스 제공 방식에 따라 구분되고 있다는 것을 알 수 있다. 웹기록물로 관리할 수 있는 웹사이트는 공개된 웹사이트 중 표면웹과 사용자 인증 등이 없는 일부 심층웹의 범위로 한정될 수 있다고 판단되는 부분이다.

〈표 2〉 웹사이트 관련 개념

구분	정의	내용
도서관법 (제2조9호)	온라인 자료	“온라인 자료”란 정보통신망(「정보통신망 이용촉진 및 정보보호 등에 관한 법률」 제2조제1항제1호의 정보통신망을 말한다. 이하 같다)을 통하여 공중 송신(「저작권법」 제2조제7호의 공중 송신을 말한다. 이하 같다)되는 자료를 말한다.
도서관법 시행령 (제13조의2)	온라인 자료	(온라인 자료의 수집) 법 제20조의2제1항에 따라 국립중앙도서관이 수집하는 온라인 자료는 전자적 형태로 작성된 웹사이트, 웹자료 등으로서 국립중앙도서관장이 제13조의3에 따른 도서관자료심의위원회 심의를 거쳐 선정하여 고시하는 자료
행정안전부 (2021)[2]	웹페이지	인터넷을 통해 텍스트, 그림, 영상, 음성 등의 정보를 제공하기 위해 만들어진 웹브라우저상의 문서
	웹사이트	특정 서비스를 위해 구성된 웹페이지의 집합체
	행정·공공 웹사이트	행정기관 등이 제공하는 정보 및 서비스를 인터넷에서 제공하는 가상공간으로, 특정 콘텐츠, 서비스 등을 위한 웹페이지들의 집합
국립국어원 (2022)	홈페이지	개인이나 단체가 월드 와이드 웹에서 볼 수 있게 만든 하이퍼텍스트 * 순화어: 누리집, 원어 homepage(영)
이지은 (2006)[3]	웹사이트	웹사이트는 인터넷의 특정 도메인을 가지며, 브라우저 기술을 이용하여 찾을 수 있는 링크된 정보자원의 조직 * 인트라넷(intranet), 익스트라넷(Extranet), 공공 웹사이트(Public website)로 구분
	인트라넷	기관 내부의 웹사이트로 공공웹사이트와 같은 프로토콜과 네트워크를 사용하지만, 내부의 이용자에게만 제한적으로 접근이 가능한 사이트
	익스트라넷	기관의 파트너로 고려된 기업이나 개인과 같이 선택된 외부인에게 접근이 허용되는 사이트
	공공웹사이트	공공웹사이트는 공공기관의 행정적 통제 아래에서 함께 링크되어 있는 전자파일 컬렉션으로, 월드와이드웹을 통해서 공공에게 접근 가능하도록 만들어져 있음

2) 행정안전부. (2021). 「웹사이트 발주자 · 관리자를 위한 행정 · 공공기관 웹사이트 구축 · 운영 가이드」, p.3

3) 이지은. (2006). ‘공공기관의 웹기록 관리방안 연구’, 한국외국어대학교 정보기록관리학과, 석사학위논문, p.8

구분	정의	내용
ISACA (2018)[4]	표면 웹 (Surface web)	검색엔진으로 접속하여 사용하는 대부분의 일반 웹사이트 * 검색엔진(구글, 빙 등), 소셜 네트워크(페이스북, 트위터, 넷플릭스 등), .com, .org와 같은 흔한 도메인 웹사이트
	심층 웹 (Deep web)	웹브라우저와 같은 일반적인 도구를 이용하여 접근할 수 있는 웹사이트로, 엄청난 양의 데이터를 제공하고 있으나 찾기 어려움 * 연구기관 및 정부의 데이터베이스, 구독하거나 비용을 지불해야 하는 웹사이트, 데이터 제공 웹사이트 등
	다크 웹 (Dark web)	기존 검색엔진으로 액세스할 수 없는 인터넷의 일부이며, 액세스하려면 특별한 익명화 소프트웨어가 필요

웹 정보자원은 쉽게 변경되거나 사라지기 쉬운 휘발적 성격을 갖고 있으나, 현대 사회에서 지식정보 공유와 커뮤니케이션을 위한 웹에 대한 의존도는 거의 절대적으로 증가하고 있다(김희정, 2010)[5]. 따라서 가치 있는 웹 자원에 대해서는 별도의 아카이빙 작업을 수행하는 것이 필요하다(김희정, 2010).

공공기관 웹기록물 관리 근거는 「공공기록물 관리에 관한 법률」에 있다. 「공공기록물 관리에 관한 법률 시행령」 제2조 제2호에서는 전자기록물을 '전자적인 형태로 작성하여 송수신 또는 저장되는 전자문서, 웹기록물 및 행정정보 데이터세트 등의 기록을 정보자료'로 규정하여 전자기록물의 범위에 웹기록물을 포함하고 있다. 동 시행령 제2조 제10호에서는 웹기록물을 '공공기관에서 운영 활동하는 웹사이트, 블로그, 소셜 네트워크 등 웹을 기반으로 생산된 기록정보자료와 웹사이트 운영 및 구축과 관련된 관리정보'로 규정한다. 전자기록물과 웹기록물의 정의에 근거 요약하면, 웹기록물은 전자기록물과 같은 정보자료로서,

4) Wanda Archy. (2018). 'Shedding Light on the Dark Web', ISACA BLOG
5) 김희정. (2010). 웹 아카이빙 인터페이스 유형 및 특성 분석, 한국기록관리학회지, 1(2), p.148

공공기관이 웹을 기반으로 생산한 자료와 운영 및 구축 관련 정보를 포함하는 개념이다.

기록은 표현을 위한 일정한 구조를 가지고 있는데, 웹사이트의 웹페이지는 면 구분이 없어져 면과 편철이라는 기록의 오래된 구조가 해체되었다(이승억, 설문원, 2021)[6]. 웹을 기록물 관점에서 접근한 차승준외(2009)의 연구에서는 웹기록물을 '표면 웹기록물'과 '심층 웹기록물'로 구분했다. '표면 웹기록물'은 사용자의 접속에 따라 동일하게 표현되는 정적인 문서의 집합이고, '심층 웹기록물'은 사용자의 요구 변경에 따라 갱신되는 동적인 웹기록물로 규정하였는데 '표면 웹이 아니며 일반적 검색엔진으로 접근되지 않는 것'이라 추가 한정하여 '동적'으로 동작하는 특징 외 '허용된 접근'에 대한 관점을 포함했다.

〈표 3〉 웹기록물 및 웹아카이빙 관련 개념

구분	정의	내용
공공기록물법 시행령 (제2조 정의)	전자기록물 (제2호)	"전자기록물"이라 함은 정보처리능력을 가진 장치에 의하여 전자적인 형태로 작성하여 송신·수신 또는 저장되는 전자문서, 웹기록물 및 행정정보 데이터세트 등의 기록정보자료
	웹기록물 (제10호)	"웹기록물"이란 공공기관에서 운영·활용하는 웹사이트·블로그·소셜네트워크서비스(SNS) 등 웹을 기반으로 생산된 기록정보자료와 웹사이트 운영 및 구축과 관련된 관리정보
국가기록원 기록물관리지침 (2019)	웹 아카이빙	World Wide Web의 일부 또는 전체를 수집하여 웹기록물 형태로 저장하는 것으로 후대의 학자, 일반대중을 위해 보존하는 것을 의미 웹기록물을 수집하기 위해 웹 크롤러(Web Crawler)라는 자동화된 툴을 사용
기록학 용어사전(2008)	웹 아카이빙	웹 자원에 대한 장기적 접근을 보장하기 위한 보존활동

6) 이승억, 설문원. (2021). 디지털 정보기술 환경에서 보존기록 평가론의 전환, 기록학연구, (67), p.65

구분	정의	내용
TTA 정보통신용어사전 (2022)	웹 기록 보관소	과거에 존재했던 특정 웹사이트를 볼 수 있는 웹사이트. Alexa 프로그램의 소유자가 처음으로 자료를 수집하였으며, 1996년부터 현재까지 자료로 100 테라바이트 메모리와 100억장의 페이지가 소요된 사이트(http://web.archive.org)이다.
국립국어원 (2022)	웹 기록 보관소	과거에 만들어졌던 특정 웹사이트의 정보를 저장해 놓고, 필요할 때 찾아볼 수 있도록 한 웹사이트
차승준, 최윤정, 이규철(2009)	표면 웹기록물	표면 웹 기록물은 접근할 때마다 동일하게 표현되는 정적인 문서의 집합으로 구성된 웹사이트로 기관소개, 연혁 등의 자료가 이에 속함
	심층 웹기록물	사용자 요구가 변경될 때마다 저장된 데이터베이스 내용도 갱신되는 웹사이트로 키워드 검색, 공지 사항, 게시물, 자료실 등이 이에 속한다. (표면 웹의 부분이 아닌 웹의 콘텐츠로 일반적 검색 엔진으로 접근되지 않는 것)

2) 선행연구 고찰

(1) 웹기록물 수집 방법

웹기록물 수집은 크게 세 가지 방법으로 구분할 수 있다. 첫 번째는 생산기관이 직접 소스와 데이터 등을 제공하는 방법, 두 번째는 기록물 관리 전문기관이 직접 복제 또는 이관하는 방법, 세 번째는 웹사이트에서 원격으로 수집하는 방법이다. 첫 번째와 두 번째 방법이 사실상 명확히 구분되지 않으므로 두 가지를 합하여 다시 구분하면 웹기록물의 수집은 '직접 이관'과 '원격 수집'으로 구분된다.

'직접 이관'의 대표적인 사례는 대통령기록관이 대통령기록물 생산기관 웹사이트를 소스 코드, 데이터베이스, 콘텐츠 등을 분리하여 이관받거나, 가상화 기술을 이용하여 운영체계를 포함하여 일체를 이관하는 방식이다. 직접 이관에 관한 한상효(2018)[7]의 보고서에서는 웹기록물이 소스 코드 등 웹 자원의 데이터 누락 없이 그대로 이관되고 그 모습 그

대로 재현 · 서비스된다는 점에서 '완전형 이관'이라 칭하였는데, 다음과 같이 '완전형 이관'의 세 가지 문제점을 지적했다. 첫 번째는 임기종료 시점(웹사이트의 서비스 종료 시점) 홈페이지만 이관되어 재현하는 한계[8], 두 번째는 고비용으로 정기적 수집이 어려워 재임 기간 중(웹사이트의 운영 기간 중) 홈페이지 변경 상황 기록화 곤란[9], 세 번째는 오프라인 이관방식의 저효율 문제를 지적했다. 이러한 지적에도, 운영당시 모습 그대로 웹기록물을 제공하고 있는 기관은 대통령기록관이 국내에서는 유일하다. 이경용(2022)[10]의 연구에서는 대통령기록관은 다년간에 걸쳐서 대통령기록물을 이관 · 정리해서 목록집이나 기술서집 등의 형태로 제시하고, 웹사이트를 통해서 다양한 기록서비스 제공을 위해 노력하고 있다고 평가했다.

대통령기록물 관리모델을 재설계한 한국무역정보통신(2018)[11]의 연구에서는 웹기록물 이관 절차를 재임 기간과 임기종료 전으로 구분하여 재설계했다. 임기종료 전 최종 이관은 현행과 같은 '완전형 이관'으로 진행하고, 재임 기간 웹사이트의 기록화는 '원격 수집' 방식으로 수행하는 복합 방식이다. '원격 수집'은 웹수집기를 이용하여 웹사이트를 원격으로 수집하는 것을 의미한다. '원격 수집'은 선행연구에서 '웹크롤링', '원격 하베스팅'으로 표현되나 같은 의미로 사용됐다.

한편 수집 대상 관점에서, 웹사이트 전체를 대상으로 하는 경우 '포괄적 아카이빙', '완전형 이관', 전체가 아닌 경우 '선택적 아카이빙', '부

7) 한상효. (2018). 영국(TNA)의 웹아카이브 동향조사, 기록인, 2018 Winter, 45, p.25
8) 한시적으로 운영기관이나 임기종료와 함께 폐지되는 기관의 웹사이트를 '완전형 이관' 방식으로 수집, 보존하는 방법의 한계로서, 최종 인수 시점의 외형만을 재현할 수 있는 한계를 의미
9) 웹사이트 이관 및 복원은 개별 웹사이트에 사용된 기술을 다룰 수 있는 소프트웨어기술자가 투입되고, 이관과 재현을 위한 하드웨어 및 상용 소프트웨어 인프라 도입이 수반되는 것을 의미
10) 이경용. (2022). 참여정부 대통령기록 연구 : 대통령 행사기록을 중심으로. 기록학연구, (71), p.133
11) 한국무역정보통신. (2018). 디지털 기반 대통령 기록관리 모델 재설계, p.223

분형 이관'이라 칭하고 있다(이성숙, 2005; 윤준희, 2008; 김희정, 2011; 한상효, 2019). 그 외 웹기록물 수집 · 보존 방법으로 기록물 관리 전문기관이 제공하는 호스팅 서비스가 검토됐다. 기록물관리기관이 웹생산기관에 가상화 기반 서버 호스팅 서비스를 제공하다가, 웹기록물 관리가 필요한 시점에 백업, 스냅샷, OVF(Open Virtualization Format) 파일[12] 생성 등의 방식으로 웹사이트와 운영환경 일체를 복제하여 웹사이트를 기록하는 것을 의미한다.

〈표 4〉 웹기록물 수집에 관한 연구

연구자	수집 방식	관련 사항
이성숙 (2005)	제출 접근	저장소에 웹사이트의 사본 혹은 스냅샷을 제출하도록 하는 방식(NARA에서 연방정부 웹사이트 아카이빙 시 사용)
	선택 접근	보존할 개별 웹사이트를 선정기준에 의해 선정하고, 웹사이트 소유자의 사용 허락을 받은 후, 미러링 SW를 이용해서 수집하는 방식
	자동 하비스팅 접근	선택 접근과 달리 하비스팅 도구를 이용하는 것으로, 대개 특정 국가(혹은 전 세계)의 모든 웹페이지를 크롤러가 링크를 따라가며 다운로드 하여 수집하는 것
유효림 (2007)	미러링 (Mirroring)	웹사이트의 소스와 DB를 완벽하게 복제하는 방법(개념적으로 웹자원을 수집하는 가장 간단한 방법)
	스냅샷 (Snapshot)	수집 로봇을 통해 자동으로 수집하는 방법으로 보편적으로 사용하는 방법 (로그, 심층웹, 업로딩/다운로드 등 로그 미획득)
	데이터베이스 아카이빙 (Database Archiving)	데이터베이스의 아카이빙을 위해 표준 자료모델과 포맷을 정의하고 각 데이터베이스 소스를 그 표준 포맷으로 변환하며 표준 접근 인터페이스로 아카이빙 된 데이터베이스를 제공하는 방법

12) OVF(Open Virtualization Foramat)은 개방형 가상화 포맷을 의미

연구자	수집 방식	관련 사항
윤준희 (2008)	완전형 이관	대상 웹기록물의 전체 이관(대통령기록물 이관방식)
	선택적 이관	일정한 기준에 따라 특정한 부문의 이관 (지정, Random 등 방식 존재)
	가상화 기반의 호스팅	가상화 기술을 이용하여 웹사이트를 운영할 수 있는 플랫폼을 호스팅하여, 이관 필요 시점에 즉시 이관 및 보존
차승준, 이규철 (2008)	컨텐츠 기반 수집	컨텐츠에 기반한, 즉 웹사이트의 기본적인 컨텐츠를 아카이빙하기 위한 방법
	이벤트 기반 수집	웹서버와 브라우저 사이에 발생된 실제적인 트랜잭션(Transaction)을 처리하는 것
정준선 (2008)	직접 전송 (Direct Transfer)	웹서버에서 직접 그 원본 데이터를 복사하여 가져오는 방식
	원격 하베스팅 (Remote Harvesting)	웹크롤러 소프트웨어를 사용하여 웹서버로부터 컨텐츠를 획득하는 방식(동적인 웹사이트 컨텐츠 수집 문제)
	데이터베이스 아카이빙(Database Archiving)	데이터베이스 기반 사이트를 아카이빙 하는 방식
	처리행위 아카이빙 (Transaction Archiving)	웹서버로부터 전달되는 컨텐츠 보다는 웹서버와 브라우저 사이에 발생하는 실제 처리행위(Transaction)를 수집하는 데 초점을 두는 것
김희정 (2011)	포괄적 아카이빙 (Extensive Archiving)	깊이는 얕지만 가능한 방대한 범위의 웹사이트를 수집하는 방식
	선택적 아카이빙 (Intensive Archiving)	대상이 되는 웹사이트 범주는 좁지만, 그 깊이를 최대한 수집하는 방식
한상효 (2018)	완전형 이관	서버 가상화를 통한 홈페이지 재구축. 소스코드 등 웹 자원의 데이터 누락 없이 웹사이트가 그대로 이관되고 그 모습 그대로 재현·서비스 (임기종료 전 최종 이관방식)
	웹 크롤링	재임 기간 중 변경 상황의 기록화(변경 상황의 기록화 주기는 연 2회로, 필요시 보존 가치가 있는 웹 기록물을 추가 수집)
TTA 정보통신용어사전 (2022)	웹 크롤러 (web crawler)	웹상의 다양한 정보를 자동으로 검색하고 색인하기 위해 검색엔진을 운영하는 사이트에서 사용하는 소프트웨어
	크롤링 (crawling)	웹사이트(web site), 하이퍼링크(hyperlink), 데이터(data), 정보자원을 자동화된 방법으로 수집, 분류, 저장하는 것

연구자	수집 방식	관련 사항
국가기록원 기록물관리지침 (2019)	온라인 원격 수집	웹 수집 로봇(Crawler)을 이용하여 원격으로 웹사이트를 수집하는 것으로, 웹페이지의 각 링크를 따라가 그 안에 포함되어 있는 텍스트, 이미지, 오디오/비디오 자료 등을 수집하는 것

(2) 웹기록물 수집 도구

선행문헌에서 공통으로 검토된 웹기록물 수집 도구는 Heritrix, HTTrack, DeepARC 등이다〈표5〉. 그중 Heritrix는 오픈 소스 기반의 웹크롤러로 약 20년간 국내외 다양한 조직에서 이용하고 있다. 웹사이트를 포괄적으로 수집·저장하는데 필요한 URI 분석 기반 자동 페이지 수집기능이 강점이다. 그러나, 최근 웹사이트는 사용자 환경의 웹브라우저에서 동작하는 스크립트 의존도가 점점 증가하고 있어 문제가 발생하고 있다. 예를 들어 페이스북이나 트위터와 같은 소셜미디어의 경우 사용자의 스크롤을 인식하여 정해진 수량의 콘텐츠가 로딩되는데 URL이나 URI의 변화가 없다. IT 기술자들은 '무한 스크롤'이라고 표현하기도 하는데, 이러한 구조는 Heritrix와 같이 URI를 이용하여 접근하는 웹사이트에서는 필요한 링크나 접속 경로를 확보할 수 없다.

Archive-it[13]은 스크립트 문제를 해결하기 위해 Heritrix를 보완하는 Umbra를 개발했다[14]. Heritrix가 URI에 접근하는 방식으로 웹페이지를 수집하면서 Umbra를 이용하여 클라이언트 측 스크립트를 처리하는 방식으로 동작한다. 이에 Heritrix 단독으로 확보할 수 없었던 URI를 감지하여 크롤링할 수 있다. 다만, Umbra는 2019년 12월 19일을 마지막으로 277번째 커밋(Commit)이 완료된 상태로 약 2년간 업데이트되지 않았다. 다양한 스크립트 기술의 등장에 따라 매일 신속하게 업데이트되는 웹

13) Archive-it은 Internet Archive에서 제공하는 웹아카이빙 서비스
14) Jillian Lohndorf. (2017). Archive-It-Crawling-Technology, Archive-It (website)

브라우저 대비, 2년의 공백은 최근 도입된 신기술을 처리하지 못하는 문제가 발생할 가능성이 있다.

〈표 5〉 웹기록물 수집 도구에 관한 연구

연구자	수집 도구	주요 설명
경기메모리 종합발전 계획 (명지대, 2016)	Heritrix	인터넷 아카이브에서 개발하여 공개 프로그램으로 배포한 리눅스 기반 웹 크롤러
	HTTrack	손쉬운 설치와 실행으로 웹사이트 수집 (Heritrix와 Wget 등에 비해 웹사이트 아카이빙 속도가 현저히 느림)
	DeepARC	심층 웹 자원을 수집이 가능한 웹 아카이빙 소프트웨어로서 심층 웹 자원 수집을 위해 관계형 데이터베이스 내용을 XML로 변환시켜주는 것을 주 목적으로 하며 자바스크립트나 플래시(flash)는 지원하지 않음
	PageVault	웹서버에서 생성되는 모든 응답(response)들의 아카이빙을 지원함 - 동적 및 정적 웹페이지에서 생성되는 모든 포맷 (HTML, XML, PDF, zip, image, sound)을 지원 - 자바스크립트, 플래시, 심층웹 수집은 미지원
	Wget	WGet(이하 WGet)은 범용 인터넷 프로토콜을 통해 대용량 파일을 다운로드하거나 웹사이트와 FTP 사이트를 전체를 쉽게 미러링하는 툴
	Archive-It	인터넷 아카이브에서 제공하는 Heritrix 기반의 웹 아카이빙 서비스
	Zotero	온라인 자원으로부터 쉽게 메타데이터를 획득하여 참조하는 툴
디지털 기반의 대통령 기록관리 재설계 (KTNET, 2018)	Heritrix	인터넷 아카이브가 개발한 오픈소스 기반의 웹크롤러
	HTTrack	웹사이트를 통째로 자신의 컴퓨터에 다운로드 하는툴
	DeepARC	프랑스 국립도서관(BnF)에서 개발하여 공개한 웹 아카이빙 소프트웨어
	Wayback Machine	인터넷 아카이브에서 운영하는 웹 아카이빙 서비스

(3) 웹기록물 보존 포맷

국제표준화 기구 및 선행연구에서 웹기록물의 보존 포맷은 WARC를 다루고 있다. 웹기록물은 ISO 28500으로 규정된 WARC 파일 포맷을 표준으로 사용하고 있다. 2009년 5월 최초 규정 후 현재 2017년 8월에 개정된 Edition 2가 최신 버전이다. 웹아카이빙 메타데이터 구조 및 요소에 관하여 연구한 오상훈 · 최영선(2009)[15]의 연구에서는 OASIS의 디지털 아카이빙 메타데이터를 '설명적 메타데이터', '구조적 메타데이터', '관리적 메타데이터', '보존적 메타데이터'로 구성했다. 이때 웹자원의 보존 포맷으로 WARC을 포함했다. 김명옥 · 이상용(2010)[16]의 연구에서는 전자기록물의 SIP, AIP, DIP와 같은 정보 패키지 재설계에 관하여 다루면서, '정보 패키지는 텍스트 기반의 전자문서에 한정되어 있고, 웹기록물, 데이터세트 등 새로운 유형의 전자기록물의 정보 패키지 규격이 핵심적 요소로 필요하다'라고 설명했다. 실제 박병주외(2010)[17]의 연구에서는 웹기록물을 새로운 유형의 포맷으로 정의하면서, '표면 웹기록물' 보존 포맷 KoSurWeb(Korea Surface Web records)과 '심층 웹기록물' 보존 포맷 KoDeWeb(Korea Deep Web records)을 설계했다. 위 포맷은 '기록물 건 메타데이터', '심층 웹 메타데이터', '기술 메타데이터', '콘텐츠'로 구성되는데, 콘텐츠 영역은 WARC를 사용하는 것으로 정의했다. 그 구성을 살펴보면, 콘텐츠에 해당하는 WARC를 중심으로 WARC를 설명하는 메타데이터가 합쳐진 포맷으로 요약된다.

그 외 단일 파일로 웹페이지를 저장하는 MHTML 포맷, SingleFile 포

15) 오상훈, 최영선. (2009). ISO14721 OASIS 참조모형을 활용한 웹아카이빙 메타데이터 구조 및 요소 정의, 정보처리학회논문지, 2009, 16, 통권 128호, p.652

16) 김명옥, 리상용. (2010). 전자기록물의 장기보존을 위한 기능요소 연구, 한국기록관리학회지, 2010, 10, p.120

17) 박병주, 차승준, 이규철. (2010). 웹기록물 보존을 위한 전자기록물 장기보존포맷 확장 설계, 한국전자거래학회지, 2010, 15, p.36

맷, RDB를 보존하는 SIARD, 웹페이지를 스크린샷으로 처리하는 PDF 등도 웹기록물의 보존에 사용할 수 있으나, Browsertrix와 같은 일부 웹수집기에서만 해당 포맷을 저장할 수 있다.

〈표 6〉 웹기록물 보존 포맷에 관한 연구 및 문헌

연구자	구분	내용
박병주, 차승준, 이규철 (2010)	KoSurWeb (Korea Surface Web)	표면 웹기록물을 보존하기 위한 문서보존포맷 (DC 메타데이터 참조 + 기술 메타데이터), XML 구성
	KoDeWeb (Korea Deep Web)	심층 웹기록물을 보존하기 위한 문서보존포맷 (SIARD 메타데이터 참조), XML 구성, ZIP64 파일 포맷으로 압축
저장포맷	WARC	웹기록물 보존포맷 국제표준(ISO 28500)
	K-WARC	WARC에 필요 메타데이터를 추가하여 개발한 한국 포맷
	SingleFile (github 공식)	SingleFile은 Chrome, Firefox, Microsoft Edge, Vivaldi, Brave, Waterfox, Yandex Browser 및 Opera와 호환되는 웹 확장 도구 ※ 하나의 HTML 파일에 완전한 웹페이지 저장 가능
	SIARD (윤성호, 2021)	스위스 연방 기록원(Swiss Federal Archives, SFA)에서 개발한 관계형 데이터베이스 보존포맷으로, 소프트웨어로부터 독립해 데이터를 보존할 수 있는 파일 포맷
	PDF (Adobe)	Adobe사가 개발한 개방형 표준 문서. ISO 32000 https://www.adobe.com/kr/acrobat/about-adobe-pdf.html

3) 이 연구의 차별성

선행연구에서 웹사이트를 웹기록물로 관리하기 위해 검토한 웹수집기는 다양한데 실제 해당 성능을 확인하지는 못하였다. 그중에는 최초 개발 이후 계속해서 업데이트 중인 '살아있는 웹수집기'가 있는가 하면, 인터넷에서 흔적도 찾을 수 없는 '사장된 웹수집기'도 있다. 기술의 발

전에 대응하여 계속해서 웹사이트를 수집할 수 있도록 업데이트되는 웹수집기와 향후 오랜 기간이 지난 후에도 열람할 수 있는 장기 보존 포맷은 웹기록물 관리를 위한 필수 요소이다. 이 연구는 웹기록물 관리에 필요한 웹수집기를 조사하고 평가하여 웹수집기를 선정하고, 일부 공공기관 웹사이트를 대상으로 원격 수집 실증 테스트를 진행하여 그 결과를 제시함으로써 현재 시점에 가장 효율적인 웹수집기를 선택할 수 있는 비교 정보를 제시하였다는 점에서 차별성을 가진다.

이하에서는 선행연구와 법령의 용어에 기반하여 웹사이트, 원격 수집, 웹수집기, 웹기록물 등의 용어를 통일하여 사용하였다.

3. 연구설계

1) 웹수집기 선정 방법

웹사이트를 원격지에서 수집하기 위해서는 반드시 웹수집기가 필요하다. 웹수집기의 성능과 기능이 원격 수집 품질과 웹기록물의 재현에 매우 큰 영향을 미친다. 웹수집기의 선정은 총 세 번의 라운드를 거쳐 선정하고, 선정한 웹수집기를 이용하여 중앙행정기관 웹사이트를 대상으로 실증 테스트를 진행한다.

라운드1은 웹수집기가 갖추어야 할 조건과 기준을 정하고 해당 기준을 충족하지 않는 웹수집기를 제외하고, 남은 웹수집기를 선정하는 방식과 필요한 핵심 기능 보유 여부 점수화하여 비교한다.

* (라운드 1) 스크리닝 & 스코어링
  - 오픈 소스 라이선스 정책
  - 국제표준 웹기록물 포맷(WARC) 저장 기능 제공

- 웹수집기 핵심 기능 보유

라운드2는 라운드1에서 선정된 웹수집기 중, 국내 및 공공부문 활용을 고려하여 개발언어, 공개소프트웨어의 개발자 참여 수, 관리 기능 보유 여부 등을 기준으로 적합하지 않은 웹수집기를 제외하고 남은 웹수집기를 통과시키는 방식으로 진행한다. 라운드2는 행정안전부 웹사이트 구축지침 및 전자정부프레임워크를 기준으로 하는 공공기관 자체 또는 기록물관리기관 등의 운영을 고려한 연속성 측면의 요소로 구성됐다.

* (라운드 2) 개발언어, 참여도, 웹수집기 관리 기능 보유 여부
  - 전자정부 프레임워크 호환 언어
  - 해당 소프트웨어의 최근 수정일, 개발자 Commit 참여 수, 관리 도구 보유 및 옵션 기능 등의 포함 여부 검토

라운드3은 라운드2에서 선정된 웹수집기를 이용하여 공공기관 웹사이트 n개에 대하여 실제 수집 테스트를 진행하고, 가장 성능 좋은 웹수집기를 최종 웹수집기로 선정하는 절차로 진행한다. 웹수집기는 자동으로 수집하는 URI 단위로 웹페이지를 저장하기 때문에 기술적으로 웹사이트 운영기관이나 원격 수집 수행기관 모두 전체 URI 수를 파악할 수 없으므로 전체 URI 대비 수집 성공 비율을 비율 또는 점수로 환산할 수 없다. 그러나, 상이한 웹수집기가 동일 시점에 동일 웹사이트를 수집하는 방식의 테스트이기 때문에 많은 수의 URI를 저장한 웹수집기의 성능이 더 좋다고 평가하는 것은 합리적이다. 또한, 동적으로 동작하는 실제 웹사이트의 용량과 웹페이지가 수집 저장된 웹기록물의 용량은 완전히 다른 개념이므로 전체 용량을 파악할 수 없으므로 URI 수와 같은 맥락에서 많은 용량을 저장한 웹수집기의 성능이 더 좋다고 판단하는 방식으로 진행한다.

* (라운드 3) 중앙행정기관 웹사이트 수집 실증 테스트
 - 웹페이지 수집 URI 수
 - 웹페이지 저장 용량

〈그림 1〉 웹수집기 선정 방법

2) 웹수집기 성능 비교 기준

라운드3에 적용할 평가 기준은 총 2가지로 기준1은 웹페이지를 수집하면서 확인할 수 있는 URI 수, 웹페이지 저장 용량으로 정하였다. 많은 웹페이지를 수집하는 경우 좋은 품질로 가정한다. 기준2는 로그 분석 정보에 정상 응답과 오류 응답을 비교하여 웹수집의 증가에 따른 정상, 비정상 로그 반영 여부를 확인하는 것으로 하였다.

〈표 7〉 웹수집기 성능 비교 기준

구분	내용
기준1	웹페이지 수집 URI 수, 웹페이지 수집 용량
기준2	로그 분석 응답(정상, 오류) 수

4. 웹수집기 선정 및 성능 비교

1) 웹수집기 선정

(1) (라운드 1) 스크리닝 & 스코어링

국내외 인터넷 웹사이트 및 소프트웨어개발자 등이 소프트웨어 소스를 공개하는 Github 등 웹사이트에서 총 23개의 웹수집기를 목록화하고 제3장에서 정한 연구설계에 따라 스크리닝과 스코어링 방법으로 라운드1을 진행하였다. 라운드1의 스크리닝 기준은 세 가지이다. 첫 번째 기준은 오픈 소스 라이선스 정책 유무이고, 두 번째 기준은 국제표준 웹기록물 보존 포맷 WARC 파일 생성 가능 유무이다. 세 번째는 웹사이트를 수집하는데 필요한 핵심 기능 보유 유무이다.

첫 번째 스크리닝 단계에서 오픈 소스 라이선스 정책이 아닌 4개의 웹수집기(Archive-It, Crawler, Octoparse, PageFreezer)를 제외하였다. 두 번째 스크리닝 기준인 WARC 파일 생성 가능 기준을 적용하여 7개 웹수집기(Craler4j, Crawjax, Gecoo, HTTrack, ItSucks, simplecrawler, WebMagic)를 제외하였다.

〈표 8〉 오픈소스 라이선스 및 저장 포맷 기준

구분		오픈소스 라이선스	저장포맷	제외여부
1	ArchiveBOX	MIT	WARC, HTML. PDF, SINGLE 등 11종	
2	Archive-It	x	ARC, WARC	제외(상용)
3	Bozzler	Apache	WARC	
4	Crawler	x	MySQL	제외(상용)
5	Crawler4j	Apache	files on disk	제외(WARC 미지원)
6	Crawjax	Apache	log file; plug-ins can do more	제외(WARC 미지원)

구분		오픈소스 라이선스	저장포맷	제외여부
7	Gecoo	MIT	자체	제외(WARC 미지원)
8	Heritrix	Apache	ARC, WARC	
9	HTTrack	GPL	files on disk	제외(WARC 미지원)
10	ItSucks	GPL	files on disk	제외(WARC 미지원)
11	NetarchviesSuite	LGPL	ARC, WARC	
12	Nutch	Apache	WARC 등 Options	
13	Octoparse	x	database, CSV, Excel, files on disk	제외(상용)
14	PageFreezer	x	web pages	제외(상용)
15	simplecrawler	BSD	files on disk	제외(WARC 미지원)
16	Squidwarc	GPLv3	WARC	
17	WAIL(Electron)	GPLv3	WARC	
18	WebMagic	Apache	files on disk	제외(WARC 미지원)
19	Webrecorder.io	Apache	WARC	
20	ArchiveWeb.page	GPLv3	WARC	
21	Browsertrix	AGPL-3.0	WACZ, WARC	
22	wget	GPL	WARC, files on disk	
23	wpull	GPL	WARC	

포괄적 수집을 위해 필요로 하는 웹수집기의 핵심 기능(링크 추적, URL 필터링, Javascript 링크 추출, Javascript 실행) 지원 여부는 해당 기관 웹사이트 또는 Github 등에서 소프트웨어개발자들이 직접 확인하였다. 위 기능을 충족하는 경우 1점, 부족한 경우 0.5점, 제공하지 않는 경우 0점으로 정하고, 핵심 기능 제공 여부를 스코어링 하였다.

최근 운영 중인 웹사이트는 대부분 Javascript가 사용되므로, 핵심 기능 중 Javascript 링크 추출 및 Javascript 실행기능이 없는 웹수집기는 (wget, wpull) 적합하지 않은 것으로 판단하여 제외하였다. 해당 과정에서 NetarchivesSuite, Squidwarc, WAIL(Electron) 웹수집기는 업데이트가 중지된 오픈 소스 프로젝트로서 지속적인 지원이 어렵다는 점에서 제외하였다. ArchiveBox는 URI 링크 추적 기능이 2단계(2 Depth)까지만 가능하므로 포괄적 수집에 적합하지 않다는 의견이 있었으나, 링크 추적

기능이 없는 것은 아니므로 조건부로 포함하였다.

〈표 9〉 웹수집기 핵심 기능 보유 여부 기준

구분		웹수집기 핵심 기능				핵심기능 계	제외여부
		링크 추적	URL 필터링	JS링크 추출	JS 실행		
1	ArchiveBOX	△(2depth)	○	○	○	3.5	조건부
3	Bozzler	○	x	○	○	3	조건부
8	Heritrix	○	○	○	○	4	충족
11	NetarchviesSuite	○	○	○	○	4	제외 (업데이트 중지)
12	Nutch	○	○	○	○	4	충족
16	Squidwarc	○	x	○	○	3	제외 (업데이트 중지)
17	WAIL(Electron)	○	○	○	○	4	
19	Webrecorder.io	○	x	○	○	3	충족
20	ArchiveWeb.page	x	x	○	○	2	충족
21	Browsertrix	○	○	○	○	4	충족
22	wget	○	○	x	x	2	제외
23	wpull	○	○	○	x	3	제외

위 기준을 적용하여 라운드1에서 ArchiveBox, Bozzler, Heritrix, Nutch, WebRecorder, ArchiveWeb Page, Browsertrix 총 7개 웹수집기를 선정하였다.

(2) (라운드 2) 개발언어, 커뮤니티 참여 수, 관리 도구

라운드1에서 선정된 7종의 웹수집기를 대상으로 전자정부 프레임워크의 연동 등을 고려하여 Java 언어 사용 여부와 최근 수정일, 소프트웨어개발자 Commit 참여 수, 관리 도구 보유 및 옵션 기능 등을 비교하였다.

먼저, 파이썬 언어를 사용하는 Bozzler, WebRecorder 웹수집기와 플러그인 방식으로만 동작하는 ArchiveWeb.page 웹수집기가 제외됐다. 라운드1을 통과한 웹수집기는 2022년 8월 최근까지도 수정되고 있었으나, ArchiveWeb.page 웹수집기는 최근 수정일과 참여 수를 확인할 수 없었다. 소프트웨어개발자가 직접 참여하여 해당 소스 코드를 수정 반영한 Commit 참여 수는 Browsertrix(3,293건), ArchiveBox(2,958건), Nutch(2,446건) 순서로 확인하였다. 관리 도구 제공 여부에서는 Nutch 웹수집기를 제외하였다.

〈표 10〉 개발언어, 커뮤니티 참여 수 등

구분		개발 언어	최근 수정일	참여 수 (Commit)	관리 도구	옵션 설정	제외 여부
1	ArchiveBOX	JAVA	2022.6.9	2,958	○ (웹화면)	-	충족
3	Bozzler	Phython	2022.7.7	144	○ (웹화면)	○ (YML)	제외 (언어)
8	Heritrix	JAVA	2022.7.9	109	○ (웹화면)	○ (XML)	충족
12	Nutch	JAVA	2022.7.28	2,446	×	○ (XML)	제외 (관리도구)
19	Webrecorder.io	Phython	2022.8.5	1,342	○ (웹화면)	×	제외 (언어)
20	ArchiveWeb.page	Plug-in	-	?-?	○ (브라우저)	×	제외 (언어)
21	Browsertrix	JAVA	2022.6.21	3,293	○ (콘솔화면)	○ (YML)	충족

라운드2에서 개발언어가 Java가 아닌 웹수집기(3종, Bozzler, Webrecorder.io, ArchiveWeb.page)와 관리 도구를 제공하지 않는 웹수집기(1종, Nutch)를 제외하고 ArchiveBox, Heritrix, Browsertrix 총 3개 웹수집기를 선정하였다.

(3) (라운드 3) 중앙행정기관 웹사이트 수집 실증 테스트

라운드3은 '특징이 다른 4개 공공기관 웹사이트'를 대상으로 웹수집 및 저장 테스트를 진행하였다. 해당 웹사이트는 다음과 같은 특징을 갖고 있다. 웹사이트1은 사진 등 이미지 콘텐츠가 많이 포함되어 있고, 웹사이트2는 텍스트 중심으로 구성되어 있다. 웹사이트3은 오랜 기간 운영되어 공지사항 및 보도자료 등 게시물 콘텐츠가 많고, 웹사이트4는 해당 기관 및 산하기관 웹사이트가 하위 도메인에 포함되어 있다. 구조상 웹사이트4가 메뉴 단계 Depth가 깊어 복잡도와 난이도가 가장 높다. 대상 웹사이트들은 소셜미디어 등에서 제공하는 무한 스크롤 방식의 웹페이지는 포함되지 않았다. 다만, 자바스크립트 방식의 〈a〉 태그는 포함되어 있다. 실증대상 웹사이트의 특징과 선정한 웹사이트는 연구진이 현재 운영 중인 중앙행정기관 웹사이트 51개 중 검토하여 정한 것으로 이미지, 텍스트의 절대적 기준에 따라 많고 적음을 정한 것은 아니다. 단, 웹사이트4는 단일 도메인에 산하기관 웹사이트가 모두 포함된 구조로는 유일하다.

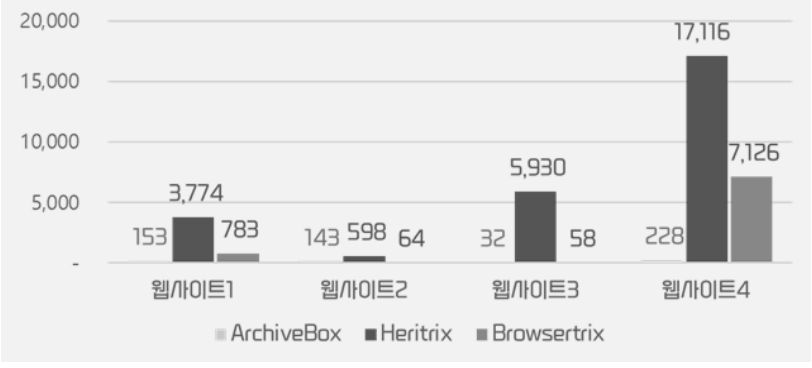
〈표 11〉 실증대상 웹사이트 특징

특징	선정한 웹사이트
사진 등 이미지 콘텐츠가 많은 웹사이트	웹사이트 1
텍스트 중심으로 작성된 웹사이트	웹사이트 2
보도자료 등 게시물 콘텐츠가 많은 웹사이트(오랜 기간 운영)	웹사이트 3
해당 기관 및 산하기관 웹사이트가 포함된 웹사이트	웹사이트 4

첫 번째 실증 테스트는 공공기관 4개 웹사이트 웹페이지 수집 URI 수 비교이다. 원격 수집을 실행한 결과 웹사이트1의 경우 ArchiveBox 153개, Heritrix 3,774개, Browsertrix 783개를 수집하였다. 웹사이트2의 경

우에는 ArchiveBox 143개, Heritrix 598개, Browsertrix 64를 수집하였다. 웹사이트2를 제외하고는 1위 Heritrix, 2위 Browsertrix, 3위 ArchiveBox 순서로 나타났다. ArchiveBox의 수집 수가 매우 적었는데, 기능상 2단계(Depth 2)를 초과하여 웹페이지를 수집하지 못하는 한계가 명확히 확인된 것으로 보인다.

〈그림 2〉 웹수집기 3종의 웹페이지 수집 URI 수(단위: 개)

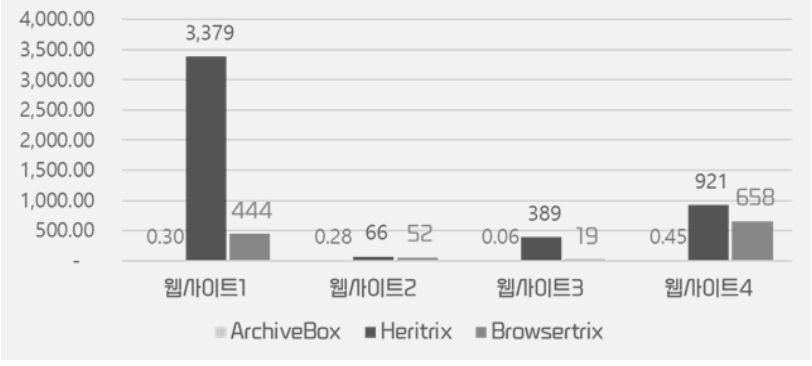


두 번째 실증 테스트는 웹수집기가 저장한 웹기록물의 용량을 비교하는 것이다. 원격 수집 실행 결과 웹사이트1의 경우 ArchiveBox[18] 0.3MB, Heritrix 3,379MB, Browsertrix 444MB를 저장하였다. 이미지가 많은 특징을 가진 웹사이트1은 Heritrix가 월등한 실적을 보였고, 텍스트가 많은 특징을 가진 웹사이트2도 Heritrix와 Browsertrix의 저장 용량이 큰 차이가 있었다. 게시물 콘텐츠가 많은 웹사이트3과 복잡도가 높은 웹사이트4의 경우에도 Heritrix가 가장 많은 용량을 저장하였다. 특이한 점은 모든 웹사이트 수집 결과에서 ArchiveBox가 저장한 웹기록물 용량이 매

18) ArchiveBox 웹수집기가 저장하는 다양한 포맷(WARC, PDF, SingleFile 등) 중 WARC 파일의 용량만을 산정하였음

우 작다는 것이다. ArchiveBox가 저장한 각 WARC 파일 용량은 2KB로 극히 작아 비정상이고, 실제 해당 WARC 파일은 열람이 불가능하여 정상적으로 저장되지 않았음을 확인하였다.

〈그림 3〉 웹수집기 3종의 웹페이지 저장 용량(단위: MByte)



특징이 다른 4개 공공기관 웹사이트를 대상으로 수집 URI 수, 저장 용량 모든 부문에서 Heritrix가 가장 좋은 수치를 보였다. 이 결과에 따르면 Heritrix는 복잡도가 높은 웹사이트3과 웹사이트4 유형을 URI를 저장하는데 비교 우위를 가진다. Heritrix는 사진 등 이미지가 많은 웹사이트1 유형을 저장하는데 탁월한 성능을 보였다.

〈표 12〉 유의미한 항목 기준 최종 비교

구분	웹수집기	수집 URL 수 (실증 테스트1)	수집 용량 (실증 테스트2)	최종 선택
공공기관 웹사이트 1	ArchiveBox	153	0.30 MByte	
	Heritrix	3,774	3,379 MByte	✔
	Browsertrix	783	444 MByte	

구분	웹수집기	수집 URL 수 (실증 테스트1)	수집 용량 (실증 테스트2)	최종 선택
공공기관 웹사이트 2	ArchiveBox	143	0.28 MByte	
	Heritrix	598	66 MByte	✔
	Browsertrix	64	52 MByte	
공공기관 웹사이트 3	ArchiveBox	32	0.06 MByte	
	Heritrix	5,390	389 MByte	✔
	Browsertrix	58	19 MByte	
공공기관 웹사이트 4	ArchiveBox	228	0.45 MByte	
	Heritrix	17,116	921 MByte	✔
	Browsertrix	7,126	658 MByte	

2) 심화 실증 테스트

(1) Heritrix 웹수집기 버전

최종 Heritrix 웹수집기가 선정됨에 따라 Heritrix 3.4(2022.7.27) 최신 버전과 Javascript 등 기술적 한계를 해결하기 위해 Archive-it이 개발한 보완 모듈 Umbra를 포함한 심화 웹수집 테스트를 진행하였다. 웹수집기1과 웹수집기2는 모든 서버 자원 조건이 같은 가상머신(VM)을 독립적으로 구성하고, 오직 보완 모듈 Umbra 미포함과 포함 여부만 다른 상태로 설정하였다. 순수한 Heritrix를 웹수집기1, 보완 모듈을 포함한 Heritrix를 웹수집기2로 정하였다.

〈표 13〉 심화 테스트 버전

구분	해당 버전	비고
웹수집기 1	Heritrix 3.4(2022.7.27)	웹수집기 단독
웹수집기 2	Heritrix 3.4.0(2022.7.27) + UMBRA	보완 모듈 포함

(2) Heritrix 웹수집 결과 비교

웹수집기1과 웹수집기2를 이용하여 공공기관 웹사이트 두 개를 대상으로 심화 테스트를 진행한 결과, Umbra 모듈을 탑재한 웹수집기가 웹사이트A 기준 1,004개의 URI를 더 수집했고, 웹사이트B 기준 2,322개를 더 수집하였다. 이때 웹사이트A는 용량이 거의 증가하지 않았고 웹사이트B는 58MB를 더 많이 저장됐다. 웹사이트를 구성하는 콘텐츠의 특성에 따라서 게시물(텍스트 정보) 등이 많이 수집된 웹사이트A는 저장 용량의 큰 차이가 없고 첨부파일 등이 포함된 웹사이트B는 URI 수집 수 증가에 따라 저장 용량도 함께 증가하였다. Umbra 모듈이 설치된 Heritrix는 순수 Heritrix 대비 수집 소요 시간이 소폭 증가하였다. 스크립트 처리에 추가 시간이 소요되기 때문이다.

〈표 14〉 웹수집기 성능(URI 수, 용량, 소요 시간) 비교

구분	공공기관 웹사이트 A		공공기관 웹사이트 B		비고
	Heritrix 3.4	Heritrix 3.4 +UMBRA	Heritrix 3.4	Heritrix 3.4 +Umbra	
수집 URI 수	8,658 개	9,572 개 (+ 1,004)	2,424 개	4,746 개 (+ 2,322)	
저장 용량	28,762 MB	28,762 MB (동일)	119 MB	177 MB (+ 58 MB)	
소요 시간	14시간 3분	16시간 26분 (+ 2시간23분)	13시간 50분	13시간 52분 (+ 2분)	

위 테스트 진행 중 생성된 로그 파일(crawl.log)을 확인하여 정상 응답과 오류 응답 수를 비교했더니 Umbra 모듈을 탑재한 웹수집기는 웹사이트A 기준으로 정상 응답 수가 894개 많았고, 웹사이트B 기준으로 2,341개 더 많았다. 그에 비해 오류 응답 수는 각 4개 증가에 머무른 것으로 Umbra 모듈의 이용이 웹페이지 URI 수집 및 정상 처리에 비례하

고 있음을 확인할 수 있다.

〈표 15〉 웹수집기 로그 분석(정상, 오류) 응답코드

구분	공공기관 웹사이트 A		공공기관 웹사이트 B		비고
	Heritrix 3.4	Heritrix 3.4 +UMBRA	Heritrix 3.4	Heritrix 3.4 +UMBRA	
정상 응답 코드	8,641 개	9,535 개 (+ 894)	2,404 개	4,745 개 (+ 2,341)	
오류 응답 코드	2 개	6 개 (+ 4개)	10 개	14 개 (+ 4개)	

5. 결론 및 향후 연구과제

1) 결론

공공부문 인터넷 웹사이트는 디자인 트랜드 또는 효율적인 정책의 전달 등을 위해 계속해서 변화한다. 수시로 변화하는 공공부문 웹사이트를 지속해서 모니터링하고 웹기록물로 보존·관리하기 위해 자동화된 방식으로 웹기록물을 수집하여 보존하는 웹수집기의 선정은 매우 중요하다. 원격 수집 방식으로 웹사이트를 완전하게 수집할 수는 없지만, 우수한 웹수집기를 사용하면 더 많은 웹페이지를 수집하고 저장할 수 있으므로 가능한 누락 없는 웹사이트를 보존할 수 있게 되기 때문이다.

인터넷상에는 다양한 웹수집기가 존재하고 사용할 수 있도록 공개되어 있으나 업무에 적합한 웹수집기를 선택하는 것은 쉽지 않다. 이 연구 결과에 따르면 보유기능 및 상대적 비교 결과에 따라 현재 시점에 원격 수집 방식의 웹기록물 관리에 적합한 웹수집기는 Heritrix로 판단된다. Java 언어로 개발되었으며, 오픈 소스 라이선스를 제공하고 많은 소프트웨어개발자가 참여하고 있으며 최근까지도 업데이트가 계속되

고 있다. 특히, 보완 모듈인 Umbra를 함께 이용하면 더 많은 웹페이지를 수집 할 수 있다. Umbra 모듈 적용은 이미지가 많은 웹사이트와 복잡도가 높은 웹사이트의 수집 결과가 향상되었고, 로그 상 수집 오류도 감소하여 양적·질적 품질 향상 효과가 있었다. 그러나, 대상이 되는 웹사이트 구축 기술은 계속 변화하고 발전하고 있으므로 Umbra와 같은 보완 모듈이 만능일 수 없다. 예를 들어 연구 과정 중 SPA(Single Page Application)[19] 기술로 구축된 웹사이트는 URI가 확보되지 않아 Heritrix 웹수집기로도 수집이 불가능하였다. 따라서 이 연구 결과가 모든 웹사이트에서 같은 결과를 가져온다고 일반화 할 수 없고, 원격수집 방식의 웹페이지 수집이 웹사이트 기록화를 완료하는 것은 아니기에 원격수집 방식과 직접 이관 방식의 복합적인 운영이 필요할 것이다.

대한민국 웹사이트는 초고속 인터넷 및 무선인터넷의 보편화로 해외 웹사이트 대비 큰 용량의 이미지가 많이 사용되고 게시판과 같은 동적 웹페이지가 많은 특징을 가진다. 그리고 웹사이트를 구현하는 새로운 기능은 앞으로도 계속 생겨날 것이다. 이 연구는 현존하는 다양한 웹수집기를 비교하였는데, 대상 웹수집기가 모두 해외에서 개발되었다. 웹 기록물 관리는 대한민국 웹사이트가 갖는 특징을 수용하고, 새롭게 등장하는 기술에 대처할 수 있어야 하므로 기존 웹수집기와 호환할 수 있는 확장성 있는 보완 모듈 확충의 개발 등 현실적인 대안이 필요하다고 판단된다.

2) 연구의 한계 및 향후 연구과제

이 연구는 모든 중앙행정기관 웹사이트를 대상으로 실증 테스트를 진행하지 못하였다. 실제 원격 수집 과정에서 20일 이상 소요되었으나

19) SPA(Single Page Application)은 서버로부터 완전한 새로운 페이지를 불러오지 않고 현재의 페이지를 동적으로 변화시켜는 웹사이트 구현 방식

수집이 종료되지 않는 웹사이트도 있어서 비교적 규모가 작은 중앙행정기관 웹사이트가 대상이 됐다. 라운드2, 라운드3 그리고 심화실증 테스트는 대국민 서비스 중인 중앙행정기관 웹사이트를 대상으로 원격 수집하였으므로 인지하지 못한 트래픽의 증가 등 확인하지 못한 다른 영향을 받았을 가능성을 배제할 수 없을 것이다.

또한, 비교 대상으로서 웹페이지 URI 수집 수, 저장 용량 등 주로 양적 기준을 적용하였는데 실제 진행한 육안검사에서 메인화면, 기관소개, 목록, 상세화면 등 4개 영역에서 품질 차이를 구분하기 힘들었다. 향후 연구에서는 WARC 파일의 정상 여부를 비교할 수 있는 기준과 비교가 다루어져야 할 것이다.

〈참고문헌〉

김명목, 리상용 (2010). 전자기록물의 장기보존을 위한 기능요소 연구, 한국기록관리학회지. 2010. 10. 101-126

김희정 (2010), 웹 아카이빙 인터페이스 유형 및 특성 분석, 한국기록관리학회지, 1(2). 147-170

박병주, 차승준, 이규철 (2010). 웹기록물 보존을 위한 전자기록물 장기보존포맷 확장 설계, 한국전자거래학회지, 2010, 15. 33-47

이경용 (2022). 참여정부 대통령기록 연구 : 대통령 행사기록을 중심으로. 기록학연구, (71), 131-167

이승억, 설문원 (2021). 디지털 정보기술 환경에서 보존기록 평가론의 전환, 기록학연구, (67), 157-97

이영남 (2018). 국가기록혁신과 기록담론, 기록학연구, (56), 49-80

이지은 (2006). 공공기관의 웹기록 관리방안 연구, 한국외국어대학교 정보기록관리학과, 석사학위논문

오상훈, 최영선 (2009). ISO14721 OASIS 참조모형을 활용한 웹아카이빙 메타데이터 구조 및 요소 정의. 정보처리학회논문지. 2009. 128(16). 651-660

차승준, 최윤정, 이규철 (2009). 공공기관 심층 웹기록물 아카이빙을 위한 매타데이터 설계, 한국전자거래학회지, 2009, 14(4). 181-193

한국무역정보통신 (2018). 디지털 기반 대통령 기록관리 모델 재설계
한상효 (2018). 영국(TNA)의 웹아카이브 동향조사. 기록인, 2018 Winter. 45. 24-29
행정안전부 (2021). 웹사이트 발주자 · 관리자를 위한 행정 · 공공기관 웹사이트 구축 · 운영 가이드

〈법률〉

공공기록물 관리에 관한 법률. [법률 제18740호, 2022. 1. 11., 일부개정]
공공기록물 관리에 관한 법률 시행령. [대통령령 제31380호, 2021. 1. 5., 타법개정]
도서관법 시행령. [대통령령 제31772호, 2021. 6. 15., 일부개정]

〈웹사이트 및 언론기사〉

Arcive-it.org, Frequently Asked Questions, website(www.archive-it.org). https://archive-it.org/blog/products-and-services/archive-it-faqs/
Jillian Lohndorf (2017), Archive-It-Crawling-Technology, Archive-It Help Center Website https://support.archive-it.org/hc/en-us/articles/115001081186-Archive-It-Crawling-Technology
Wanda Archy (2018), Shedding Light on the Dark Web, ISACA BLOG. https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2018/shedding-light-on-the-dark-web