복합적 웹 아카이빙 정책에 관한 고찰*

- 프랑스국립도서관의 사례를 중심으로 -

A Study of Combined Web Archiving Policy: BnF's Three Layers Web Archiving Strategy

김 유 승 (You-Seung Kim)**

목 차

초 록

본 연구는 웹 정보자원이 우리 삶의 전 영역에서 지니는 가치와 중요성에 비해, 이를 지키고 보존하려는 제도적 환경과 노력이 미흡하다는 인식 아래, 우리의 웹 정보자원 수집·보존정책의 발전 방향에 대한 논의를 목적으로 한다. 이를 위해 범위, 방법, 품질이라는 세 가지 측면에서 웹 아카이빙의 특성을 분석하고, 비교 모델로서 도식화한다. 이러한 분석에 기반 하여 각국의 웹 아카이빙 사례를 7가지 유형으로 나누어 각각의 장점과 단점을 진단하고, 최적의 웹 아카이빙 접근법으로서의 복합적 아카이빙 정책에 대해 논의한다. 이에 복합적 웹 아카이빙 정책을 채택하고 있는 프랑스 국립도서관의 웹 아카이빙 사례를 심층적으로 살펴본 후, 우리나라의 국가적 웹 아카이빙 프로젝트 OASIS를 비판적으로 분석한다. 결론으로, OASIS 프로젝트, 나아가 우리나라 웹 아카이빙 정책 발전을 위한 두 가지 방안을 제시한다.

ABSTRACT

This study aims at discussing development of web archiving policies in South Korea. The study is based on the understanding of that the institutional environment and efforts for keeping Web information resources are insufficient, compared to the value and importance of them. For the study, Web archiving practices are analyzed into three aspects: scope, method, and quality. Futhermore, they are graphically schematized as a comparative analysis model. Based on the model, the study classifies national Web archiving practices into seven unique types and diagnoses their cons and pros. In this context, a combined Web archiving policy is discussed as an optimal Web archiving approach. As a case study, the France National Library's Web archiving is discussed in depth and the Korean National Library's Web archiving project, OASIS, is critically analyzed. As a result, the study proposes two alternative plans for the development of Web archiving policy in South Korea.

키워드: 웹 아카이빙, 복합적 아카이빙, 선택적 아카이빙, 디지털 납본, OASIS
Web Archiving, Combined Archiving, Selective Archiving, Digital Deposit, Oasis

---

* 이 연구는 2008학년도 중앙대학교 학술연구비 지원에 의해 이루어진 것임.
** 중앙대학교 문헌정보학과 조교수(kimyus@cau.ac.kr)
논문접수일자: 2008년 11월 11일 최초심사일자: 2008년 11월 27일 게재확정일자: 2008년 12월 9일

1. 들어가는 글

모든 길은 웹으로 통한다. 정부 행정에서부터, 상거래, 학습, 그리고 공적 담론 형성과 여가생활에 이르기까지 정치, 사회, 문화 활동의 무대가 웹으로 확장되고 있다. 이러한 맥락에서 우리의 삶을 기록하고 보존하며 후세에 이용가능하게 하는 아카이브의 영역에서 웹 기반 정보자원의 중요성 또한 날로 커져가고 있다. 하지만 웹 정보자원의 기록과 보존이라는 과제는 간단치 않다. 분산성, 휘발성 등 웹 기반 정보자원의 태생적 특성[1]에 기인한 문제들에서부터, 보존적 측면에서의 진본성 · 무결성과 관련된 기술적 문제, 그리고 웹 정보자원의 수집활동에 따른 법적 문제들에 이르기까지 숱한 문제들이 산적해 있다. 이에, 다수의 나라들은 웹 아카이빙에 대한 기술적 표준 개발과 제도적 근거 마련에 많은 노력을 기울이고 있다. 호주의 PANDORA,[2] 영국의 UKWAC,[3] 미국 국회도서관의 Minerva,[4] 프랑스국립도서관 등이 그 대표적인 예이며, 우리나라에서는 국립중앙도서관의 OASIS(Online Archiving and Searching Internet Sources)가 국가적 웹 정보자원의 보존과 수집에 나서고 있다.

그러나 웹의 전 지구적 성격에도 불구하고, 각 나라, 각 기관들의 웹 정보자원 수집과 보존에 대한 범위와 방식은 동일하지 않다. 근본적으로 수집의 대상을 결정하는 가치 기준과 지리적 범주의 설정이 다를 뿐만 아니라, 수집활동의 방식과 그에 따라 적용하는 기술적 해법 또한 상이하다. 지정된 주제를 대상으로 선택적이며 제한적인 웹 정보자원을 수집하는 경우에서부터, 가능한 포괄적인 영역에서 자동화 기술을 사용하여 웹 아카이빙에 나서고 있는 나라들까지 다양한 모습으로 나타나고 있다. 그 중 후자에 속하는 일부 국가들은 웹 아카이빙의 제도적 근거로서 웹 정보자원의 납본을 법제화하고 있다.[5] 이들 나라들의 납본법의 구체적인 대상, 방법, 절차, 보상체계들은 매우 다양함에도 불구하고, 납본의 대상을 기존의 오프라인 매체에서 온라인 정보자원으로 확대했다는 점과 그 권한과 책임을 국립도서관에 위임하고 있다는 점에서 예외 없이 일치하고 있다.

우리나라의 『도서관법』 제2조는 도서관이 다루고 제공하는 "도서관자료"를 "도서관이 수집 · 정리 · 보존하는 자료로서 인쇄자료, 필사자료, 시청각자료, 마이크로형태자료, 전자자료 그 밖에 장애인을 위한 특수자료 등 지식정보자원 전달을 목적으로 정보가 축적된 모든 매체"로 규정함으로서 국립중앙도서관이 국가적 웹 아카이빙의 주체로 나설 수 있는 법적 근거를 제공하고 있다(김유승 2007, 20). 하지만

1) 웹 페이지의 수명에 대한 정확한 통계를 얻기는 쉽지 않다. 하지만 잘 알려진 여러 연구를 통해 일반적인 웹페이지의 수명은 평균 75일이고 30%가 조금 넘는 URL들만이 1년 이상 유지되는 것으로 조사된 바 있다(Lawrence 2001; Koehler 2004).

2) Preserving and Accessing Networked Documentary Resources of Australia. [citied 2008. 10. 3]. 〈http://pandora.nla.gov.au〉.

3) UK Web Archiving Consortium. [citied 2008. 10. 3]. 〈http://www.webarchive.org.uk/〉.

4) 〈http://www.loc.gov/minerva/〉 [cited 2008. 9. 25].

5) 웹 정보자원의 납본을 법제화한 나라로는 캐나다, 덴마크, 프랑스, 독일, 노르웨이, 남아프리카공화국, 스웨덴, 영국, 아이슬랜드 등이 있다(Henriksen 2001; National Library of Australia 2008a; 곽승진 외 2008).

우리의 납본법 제도는 아직 웹 정보자원을 납본 대상으로 규정하고 있지 않다.[6] 『도서관법 시행령』 제13조는 납본 대상으로 콤팩트디스크, 디지털 비디오 등 유형물과 함께 "출판환경의 변화에 따라 새로운 형태로 발간되는 기록물로서 문화관광부장관이 인정하는 자료"를 명시하고 있다.[7] 여기서 "새로운 형태로 발간되는 기록물"이 웹 기반 정보자원을 포괄하느냐는 해석의 여지가 있으나, 일반적으로는 그렇지 아니한 것으로 인식되고 있다.

본 연구는 웹 정보자원이 우리 삶의 전 영역에서 지니는 가치와 중요성에 비해, 이를 지키고 보존하려는 우리의 제도적 환경과 노력이 미흡하다는 인식에서 출발하여, 웹 아카이빙 정책의 문제를 분석하고 대안을 제시하는 것을 목적으로 한다. 이를 위해 웹 아카이빙의 특성을 분석하고, 기존 웹 아카이빙을 유형별로 구분하여 논한다. 특히, 웹 아카이빙 방법에서의 이분법적 선택을 너머선 최적의 정책적 대안으로 제시되고 있는 복합적 웹 아카이빙을 프랑스국립도서관의 사례를 통해 분석하고, 이를 통해 국립중앙도서관 웹 아카이빙 시스템 OASIS의 발전 방향에 대한 논의를 이끌어내고자 한다.

2. 선행연구

우리나라에서는 선도적 해외 웹 아카이빙 프로젝트들의 성과와 문제점들을 검토 연구한 서혜란(2004)의 『웹 아카이빙의 성과와 과제』를 시작으로 하여, 웹 정보자원의 보존에 관한 학술적 연구가 활발히 전개되어 왔다. 대표적인 연구 성과들로는 이소연(2004)의 『디지털유산의 장기적 보존: 국가정책수립을 위한 제언』, 송병호(2005)의 『진본성 확보를 위한 전자기록물 관리 방안』, 임진희(2006)의 『OAIS 정보모델과 기록 AIP』 등을 들 수 있는데, 대부분의 연구들은 웹 정보자원의 보존이라는 포괄적인 측면보다는 OAIS 참조모형을 중심으로 한 연구에 집중해왔다. 김유승(2007)의 『웹 아카이빙의 법제도적 문제에 대한 고찰』에서 웹 정보자원의 포괄적 수집방법의 채택을 포함한 웹 아카이빙 정책 전반에 대한 문제제기가 있었으나, 그 실효성과 타당성에 대한 구체적 논의로까지 발전시키지 못한 아쉬움이 있다. 곽승진 외(2008)는 『디지털자료 납본에 대한 보상 체계 연구』에서 주요국가들의 온라인 정보자원을 포함한 디지털 자료의 납본 현황을 조사하고, 보상 기준을 제시하였으나, 음악 및 전자책을 비롯한 다양한 디지털 정보자원 전반의 납본 및 보상 정책에 집중한 바, 웹 정보자원 수집정책과 관련된 문제점들은 구체적으로 다루지 않고 있다.

3. 웹 아카이빙의 특성별 비교·분석

각국, 각 기관이 시행하고 있는 웹 아카이빙은 사회·정치적 이벤트 또는 사건을 주제로

6) 2007년 9월 입법 발의된 '디지털자료 납본 및 이용에 관한 법률안'은 저작권과 납본보상금 등의 문제로 인하여 출판계의 강한 반발에 직면하여 법 제정의 시기와 가능성이 불투명하다.
7) 『도서관법 시행령』 제13호 제8호. [일부개정 2008. 6. 5 대통령령 제20797호].

한 웹 정보자원의 컬렉션에서부터 전 세계 웹 전체를 대상으로 한 아카이빙에 이르기까지 매우 다양한 모습들로 전개되고 있다. 이러한 차이들은 무엇보다도 각 아카이브들이 웹 정보자원 수집의 범위를 어떻게 설정하느냐에 따라, 그리고 어떠한 방법을 채택하느냐에 따라 크게 구분지어지며, 이와 연계되어 수집되는 정보자원의 품질에도 큰 차이들이 생겨난다. 다음에서는 범위, 방법, 품질이라는 세 가지 측면에서 웹 아카이빙을 비교·분석해보고자 한다.

3.1 범위

전통적인 납본의 개념에서 수집의 대상은 인쇄물, 사진, 음반 등 물리적 형태를 가진 공적 영역의 간행물로서 한 국가의 영토 내에서 출판되고 유통되는 것을 전제 조건으로 하였다(Lasfargues, Oury and Wendland 2008). 다시 말해, 수집의 범위에서 사적 혹은 영리조직 영역은 철저히 제외되었으며, 한 국가의 사법권이 미치는 곳으로 한정되어 왔다. 하지만 웹 정보자원의 수집에는 이러한 수집 범위의 전통적 개념을 적용하기가 쉽지 않다. 웹은 공적 커뮤니케이션과 사적 커뮤니케이션의 영역을 뒤섞어 경계를 모호하게 하는 태생적 특성을 지니고 있을 뿐만 아니라, 웹 공간에서 일어나는 정보의 흐름과 이용자들의 행위는 일상적으로 한 국가의 사법권을 벗어나 이루어지기 때문에 사법권이라는 지리적 한계를 바탕으로 한 기존의 제도들을 적용하기 매우 어렵다(Johnson and Post 1996).

그러므로 웹 정보자원 수집의 지리적 범위 선정에는 서로 다른 기준이 적용될 수 있다. 현재 각국에서 수행되고 있는 웹 아카이빙의 지리적 범위 설정은 다음의 두 가지 유형으로 크게 나뉜다.

첫째는 한 국가의 사법권이 행사되는 지리적 경계를 기준으로 두는 전통적인 방식으로, 프랑스 국립도서관의 웹 정보자원 수집지침이 대표적인 예다. 이 경우, 자국의 최상위국가도메인(nTLD) 아래 등록된 모든 사이트들과 그 외에 자국 영토 내에서 제작되고 호스팅 되어 있는 사이트들 수집 대상으로 한다. 둘째는 지리적 제약에 상관없이 한 국가의 정치, 사회, 문화를 망라한 모든 것을 수집 대상으로 하는 것으로 스위스 국립도서관이 이 정책을 채택하고 있다(BnF 2006). 국립중앙도서관의 OASIS도 '자원수집지침'에서 "한국과 관련된" 중요 정보자원을 수집대상으로 함으로써 수집 대상이 국내에 제한되고 있지 않음을 밝히고 있다. –단, 해외 자원일 경우, 한국 사람에 의해 써지거나 한국을 주제로 다룬 것 등으로 제한한다(OASIS 2006a).

또 다른 측면에서 웹 아카이빙의 수집 범주는 도메인 중심과 주제 중심으로 나뉜다. 도메인 중심 웹 아카이브는 콘텐츠에 의해 주도되는 것이 아니라 콘텐츠의 위치에 의해 좌우된다. 이러한 접근법은 일반 최상위도메인과 국가 최상위도메인을 중심으로 한 광범위한 영역을 포괄한다. 반면, 주제 중심 아카이빙은 선거 사이트를 비롯한 정치, 사회, 문화적으로 중요성을 지니는 선택된 웹 콘텐츠들에 대한 아카이브를 수행했다(Cruse, Eckman and Kunze 2003; Lyle 2004). 미국 국회도서관의 웹 아카이빙 프로젝트 Minerva가 그 대표적인 예다.

이러한 맥락에서 전자를 전체 도메인 접근법

(whole domain approach), 후자를 선택적 접근법(selective approach)라고 한다.

3.2 방법

수집 대상의 범위 설정은 수집의 방법과 결합하여 더욱 다양한 수집 형태로 나타난다. '전체 도메인 접근 방식'이 광범위한 영역의 수집에 적합한 자동 하베스트 방식을 채택하는 반면, '선택적 접근법'에서는, 자동화 방식의 보조적 사용에도 불구하고, 일정 단계의 매뉴얼 방식 채택이 일반적이다. 따라서 선택적 아카이빙은 전체 도메인 아카이빙보다 상대적으로 자원 집중적이며, 고비용이다.[8)]

이론적으로, 전체 도메인 하베스팅 접근법의 명백한 장점은 모든 도메인이 로봇에 의해 일정 주기 간격으로 자동 수집된다는 것이다. 따라서 인적 자원의 간섭은 최소화되며 수집되는 아이템 당 인건비는 상대적으로 낮다. 이론적으로는 운영자 한 명이 로봇을 운영하여 링크 탐지와 추적을 통해 수백만 개의 사이트를 발견하고 다운로드 할 수 있다. 다시 말해, 가장 광범위한 영역의 도메인을 미래 연구자들에게 이용가능해줄 뿐만 아니라, 웹 정보자원을 다른 문헌들이 보유한 링크와 함께 광범위한 맥락에서 관찰될 수 있게 한다.

하지만 실제는 이와 큰 차이가 있다. 자동화 방법은 분명한 한계를 지니고 있다. 첫째 동적(dynamic) 웹 정보자원 혹은 숨겨진 웹이라 불리는 심층 웹(deep web) 정보자원의 경우, 자동화 방법으로 수집이 불가능하다. 패스워드 혹은 접근 장벽을 채용한 상업 사이트는 로봇이 접근할 수 없다(Masanés 2002).[9)] 둘째, 자동화 방식의 또 다른 단점은 새로운 사이트를 발견하는 데 요구되는 지연이다. 모든 도메인을 대상으로 하는 하베스팅은 컴퓨터 시간과 저장공간을 요구한다. 새로운 사이트를 찾기 위한 링크의 사용은 전 지구적 크롤링에서 긴 프로세스가 될 수 있다. 수개월 간격으로 하베스팅한다고 하더라도 그 사이에 어떤 정보자원이든 나타났다가는 사라질 수 있으며, 그 사이에 변화를 모두 담아낼 수도 없다. 크롤러가 특정 이벤트와 관련된 한시적 사이트에 도달하는 것을 목적으로 한다면, 이 시간적 지연은 관련된 정보자원의 위치를 찾아내고 아카이브하기에 너무 길 수 있다. 셋째, 수집물의 품질과 완결성에 관한 비판이다. 거대한 양의 정보자원을 다루기 때문에 품질에 대한 점검은 아주 작은 표본을 통해서만 이루어진다.[10)] 다섯째, 인적 자원의 비용은 선택적 정보자원을 수집하는 매뉴얼 방식에 비해 낮지만, 데이터의 처리와 저장, 그리고 중장기 보존이란 측면에서 상당한 비용이 필

8) 이러한 이유로 웹 로봇을 이용한 자동화 방식이 많이 채택되고 있다. 각국의 8개 웹 아카이빙 기관의 수집방법을 조사한 한 연구에 의하면 자동 하베스팅 방식을 채택한 기관이 4개, 다른 방법과 자동화 방식을 채택하고 있는 기관이 1개, 그리고 나머지 3곳만이 선택적 접근 방식을 채택하고 있었다(Day 2003, 18). 하지만 주제 중심의 선택적 수집 과정에서도 로봇의 사용은 일반화되어 있다.

9) 주제 중심의 선택적 접근법에서도 심층 웹의 수집을 위해서는 정보자원의 저작권와의 개별적 계약과 적절한 기술적 해법의 적용이 필요하다. 또한, 주제 크롤링(subject crawling)의 정교함과 효율성이 날로 발전되고 있지만, 여전히 최고의 정보자원으로의 직접 연결을 제공하는 전문가들의 네트워크와 비교될 수 없다(Bergmark 2002; Bergmark, Lagoze and Sbityakov 2002).

10) Phillip(2003, 10)의 연구는 40%의 하베스트 타이틀이 불완전하거나 결함이 있는 것으로 보고하고 있다.

요하다. 이는 선택적 매뉴얼 수집 방식을 통해 수집된 대부분의 정보자원들은 품질이 평가되고 목록으로 작성되어 접근과 이용이 편리하다는 점과 대조적이다(Phillips 2005, 60).

하지만 선택적 접근법도 많은 문제를 가지고 있다. 무엇보다도 이 방식은 노동집약적이다. 한 번에 아카이브 될 수 있는 자원의 양은 가용한 인적자원의 수에 크게 좌우되며 비례한다. 또한 정보자원의 맥락을 고려하지 못하며, 곧잘 링크된 다른 리소스들을 포함하지 않는다. 따라서 일부 맥락적 의미가 사라질 위험이 높다. 이러한 측면에서, 선택적 아카이빙은 여전히 출판물이라는 개념에 대한 전통적 인식에 크게 기반하고 있는 것으로 보인다. 선택적 아카이빙의 범위는 한 국가 도메인의 자원량에 비해 지극히 제한적이다. 이러한 자원들의 대부분이 아무런 미래적 연구 가치를 지니지 않을 수도 있지만, 그 와중에 가치를 지닌 정보자원이 함께 사라질 것은 분명하다. 우리는 우리가 무엇을 잃고 있는지조차 알 수가 없다. 바로 이 점이 선택적 아카이빙을 채택하고 있는 수집 기관들이 자원의 미래적 가치에 대한 주관적 판단이라는 비판에서 자유롭지 못한 이유다. 도서관의 사서들에게 이와 유사한 장서개발결정은 낯선 업무가 아니다. 하지만 인쇄물 환경은 그러한 결정을 내리기에 좀 더 구조화, 안정화, 예측 가능한 확고한 환경이다. 웹은 지금 이 순간에도 발전을 거듭하고 있다. 불과 10여 년 전 누구도 웹의 잠재력을 정확히 예측해내지 못하였듯이 다시 10여 년 후 웹 정보자원의 잠재력이 어떻게 발전하고, 접근 · 이용 · 적용될지 예견하기란 쉬운 일이 아니다. 이러한 맥락에서, 현재의 전문가적 경험과 판단에 근거한 웹 정보자원의 미래적 가치 예단은 소중한 정보자원의 대량 손실이라는 위험을 불가피하게 전제하고 있다. 이러한 단점들은 쉽게 간과될 수 없는 것들이다.

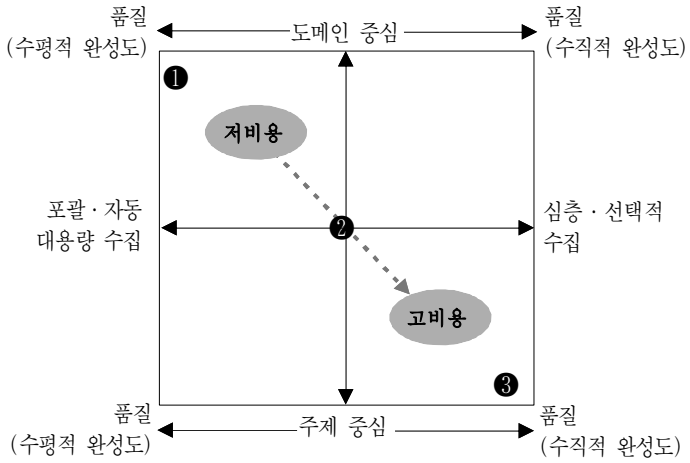
3.3 품질

웹 아카이빙의 품질은 수집된 정보자원이 원래의 완결된 형태와 기능을 유지하고 있는가에 관한 문제이다. 따라서 아카이빙 된 정보자원은 첫째 정보자원 자체의 완성도, 둘째 내비게이션과 사용자의 상호작용과 관련하여 사이트의 원래 형태를 제공하고 있는가에 의해 그 품질이 규정될 수 있다. 완성도는 발견된 관련 진입점(일반적으로 사이트의 홈페이지)의 수에 의해 수평적으로 그리고 이 진입점으로부터 발견된 관련 링크 노드의 수에 의해 수직적으로 측정될 수 있다. 이상적으로 아카이브는 수평적으로 그리고 수직적으로 완결되어야 한다. 하지만, 현실적으로는 두 가지 품질 중 어느 하나에 더 큰 무게 중심을 두는 정책 사례가 많다. 수직적 완성도보다 수평적 완성도를 선호하는 경우, 이를 포괄적(extensive) 아카이빙이라고 한다. 반대로 수직적 완성도를 우선시하는 경우를 집중적(intensive) 아카이빙이라고 부른다(Masanès 2005b, 77-78; Kim and Lee 2007, 146-148).

3.4 특성별 비교 모델

앞서 논의한 바와 같이 웹 아카이빙의 범위, 방법, 품질을 중심으로 한 분석을 도식화 하면 〈그림 1〉과 같다.

세로축은 수집의 범주에 있어서 도메인 중심



〈그림 1〉 웹 아카이빙의 특성별 비교 모델

과 주제 중심의 스펙트럼을 나타내고 있고, 가로축은 수집의 방법적 측면에서 자동화 기술을 이용한 포괄적 · 대용량 수집 방법에서부터 심층적 · 선택적 수집의 스펙트럼을 보여주고 있다. 품질에 있어서는 수집의 범주와 상관없이 포괄적 수집과 선택적 수집이 각각 수평적 완성도와 수직적 완성도에 중점을 두고 있음을 도식화하고 있다. 마지막으로, 비용 측면에서는 도메인 중심의 포괄적 대용량 수집이 상대적으로 낮은 비용인 반면, 선택적 수집이 강화될수록, 주제 중심으로 수집의 범주가 이동할수록 상대적으로 높은 비용인 것을 나타내고 있다.

〈그림 1〉의 왼쪽 상단의 ❶ 영역에 속하는 대표적인 웹 아카이브로는 스웨덴 왕립도서관의 웹 아카이브 프로젝트인 Kulturarw3[11]와 미국의 Internet Archive[12]를 들 수 있다. 이 두 웹 아카이브 모두 수집로봇을 이용한 포괄적 수집원칙을 가지고 있다. 반면, 선택적 주제 중심의 아카이빙을 수행하는 미국 국회도서관의 Minerva는 오른쪽 하단 ❸ 영역에 속한다.

하지만 주제 중심과 도메인 중심 또는 포괄적 아카이빙과 집중적 아카이빙이라는 이분법적 선택만이 존재하는 것은 아니다. 심층 웹 정보자원에 대한 수집 기술은 날로 발전하고 있으며,[13] 위의 여러 접근법의 복합적 운용도 효과적인 정책 대안이 될 수 있기 때문이다. 수집물의 높은 품질과 완결성이 아카이빙의 핵심이라고 할 때, 어느 한 방식만을 고집하는 것보다 이를 복합적으로 운영하는 것이 웹 아카이브 발전에 큰 도움이 될 것으로 보인다 – 이러한 복합적 방식을 채택한 대표적인 나라가 프랑스이다. 〈그림 1〉에서 ❷ 영역에 위치한다.

11) 〈http://www.kb.se/english/find/internet/websites/〉 [citied 2008. 10. 4].
12) 〈http://www.archive.org〉 [citied 2008. 10. 4].
13) 한 연구에 따르면, 다섯 레벨 깊이의 크롤링으로 한 웹사이트의 유용한 콘텐츠의 90%에 다다르기에 충분한 것으로 나타났다(Baeza-Yates and Castillo 2004).

4. 웹 아카이빙 사례의 유형별 분석

앞서 논의하였듯이 웹 아카이빙의 유형은 수집 범주를 기준으로 '선택적 아카이빙,' '전체 도메인 아카이빙,' 그리고 이 두 가지 방법을 병행하는 '복합적 아카이빙'으로 크게 구분된다. 선택적 아카이빙은 수집 대상에 따라 다시 '주제별 웹 자원의 선택적 아카이빙,' '정적 웹 자원의 선택적 아카이빙' 그리고 '정적·동적 웹 자원의 선택적 아카이빙'으로 세분되며, 전체 도메인 아카이빙은 '단일 국가 전체 도메인을 대상으로 한 아카이빙'과 '전 세계 전체 도메인을 대상으로 한 아카이빙'으로 나뉜다. Margaret Phillips는[14] 이러한 유형 분류에 '선택된 상업 정보생산자와의 협동적 계약에 근거한 아카이빙'을(Phillips 2003, 8; 2005, 60), 그리고 호주 국립도서관의 PADI는[15] '납본 기반 아카이빙'을 추가하였다(National Library of Australia 2008b; 곽승진 외 2008). 이러한 분류틀에 기반하여 본 연구는 웹 아카이빙을 3가지 대분류 유형과 7가지 세부 유형으로 구분하고, 이를 웹 정보자원 납본제도 도입여부와 비교하였다(표 1 참조).

4.1 선택적 아카이빙

선택적 아카이빙의 첫 번째 유형인 '주제별 웹 정보자원의 선택적 아카이빙'의 대표적인 예로는 미국 국회도서관 웹 아카이브(The Library of Congress Web Archives, 이하 LCWA)를 들 수 있다. LCWA는 2000년 35개 선정 웹사이트를 대상으로 한 아카이빙 실험인 Minerva 프

〈표 1〉 웹 아카이빙의 유형

대분류 유형	세부 유형	국가/주체	*	근거법률
선택적 아카이빙	주제별 웹 자원의 선택적 아카이빙	미국 국회도서관	X	—
	정적 웹 자원의 선택적 아카이빙	캐나다 국립도서관	O	Library and Archives of Canada Act, 2004
	정적·동적 웹 자원의 선택적 아카이빙	호주 국립도서관	X	National Library Act, 1960
	선택된 상업 정보생산자와의 협동적 계약에 근거한 아카이빙	네덜란드 국립도서관	X	—
복합적 아카이빙	선택적, 포괄적 아카이빙의 병행	프랑스 국립도서관	O	Law on Authors' Rights and Related Rights in the Information Society
		노르웨이 국립도서관	O	The Norwegian Legal Deposit Act, 1989
전체 도메인 아카이빙	단일 국가 전체 도메인을 대상으로 한 아카이빙	스웨덴 국립도서관	O	The Legal Deposit Act, 1993
		아이스란드 국립도서관	O	The Legal Deposit Law, 2002
	웹 전체를 대상으로 한 아카이빙	미국 Internet Archive	X	—

* 웹 정보자원 납본제도의 도입 여부

14) 호주국립도서관 디지털아카이빙 책임자.
15) Perserving Access to Digital Information, National Library of Australia.

로젝트로 시작되었으며, 주제 전문가에 의해 선정된 선거, 테러, 전쟁 등 지정 주제에 대한 웹 기반 정보자원을 평가, 선정, 수집, 목록화, 보존하여 접근을 제공하는 것을 목적으로 한다.[16] 2008년 현재 17개 웹 아카이브 컬렉션을 완성하였고, 3천여 개 웹 사이트를 포괄하는 네 개의 웹 아카이빙 컬렉션을 수립하고 있다.[17]

둘째 유형인 '정적 웹 자원의 선택적 아카이빙'에서는 상호작용적 혹은 동적 요소를 포함하고 있지 않은 인쇄 간행물과 유사한 웹 정보자원이 선택적으로 수집되는데, 이 그룹에 속하는 대표적인 나라가 캐나다다. 캐나다는 일찍이 1994년 최초의 웹 아카이빙 프로젝트로 알려진 Electronic Publication Pilot Project를 통해 자발적 납본 형식으로 온라인 간행물을 수집해왔으며, 2004년 Library and Archives of Canada Act의 제정을 통해 온라인 간행물에 대한 법적 납본 제도를 도입하였다. 하지만 캐나다 국립도서관은 납본의 대상이 되는 전자 간행물의 범위를 상대적으로 좁게 설정하였는데, 온라인 데이터베이스, FTP, BBS 등을 제외한 전통적 간행물로서의 특징을 갖춘 온라인 정보자원만이 납본의 범주에 포함되고 있다(Electronic Publication Pilot Project team and Electronic Collections Committee, 1996, 8-9; 김유승 2007, 14).

세 번째 방식은 정적 웹 자원 이외에 동적 웹 자원까지를 수집 대상에 두는 것으로, 집중적인 인적 자원의 투입이 필요하다는 점에서는 앞서 방식들과 동일하다. 호주의 국가 웹 아카이브 PANDORA가 이 방식을 채택하고 있다. PANDORA는 호주국립도서관을 비롯한, 1개의 준주 도서관과 6개의 주 도서관 등 10개 파트너 기관의 협력을 기반으로 PANDAS (PANDORA Digital Archiving System)에 의해 운영되고 있다(Koerbin 2004).

그 외에 선택된 상업 정보생산자 혹은 출판업자와의 협동적 계약에 근거한 아카이빙 정책은 네덜란드 국립도서관이 채택하고 있다(Phillips 2005, 60). 이는 전 세계 과학 출판물의 30% 정도가 네덜란드 내에서 생산되고 있는 특수한 상황과 맞물려 있다. 하지만, 이러한 특정 환경이 아니라 하더라도, '리치 웹(Rich Web)' 정보자원의 많은 부분을 포함하며 수집 로봇이 미치지 못하는 심층 웹의 수집을 위한 적극적인 해법으로서 중요성이 인식되고 있다(National Library of the Netherlands 2004).

4.2 전체 도메인 아카이빙

전체 도메인 아카이빙에서 가장 보편적인 접근법은 단일 국가 전체 도메인에 대한 하베스트를 기반으로 하는 아카이빙 방법이다. 이는 앞서 선택적 방식과는 크게 달리, 일반적으로 각국 국립도서관들이 자국의 전체 웹 도메인을 수집 범위로 하여 자동 하베스트를 수행한다. 이를 위해 하베스팅 로봇을 사용하며 인간의 간섭

16) 〈http://www.loc.gov/minerva/〉 [citied 2008. 10. 2].

17) 현재 LCWA 사이트에서 이용 가능한 컬렉션은 다음과 같다: United States 107th Congress Web Archive 2001; United States 108th Congress Web Archive; Crisis in Darfur, Sudan, Web Archive, 2006; Library of Congress Manuscript Division Archive of Organizational Web Sites; Papal Transition 2005 Web Archive; September 11, 2001 Web Archive; United States Election 2000 Web Archive; United States Election 2002 Web Archive; Visual Image Web Sites Archive.

은 최소화된다. 이 방식은 특정 도메인 상의 자원만을 수집하는 것이 아니라, 일반 도메인에서도 주제 혹은 출처를 식별하여 수집 범위에 둔다. 스웨덴, 아이슬란드 등이 이러한 방식을 채택한 대표적인 나라들이다. 스웨덴은 이미 1996년부터 앞서 언급한 Kulturarw3 프로젝트를 통해 가능한 많은 정보의 수집을 목적으로 하는 포괄적 웹 아카이빙 정책을 채택한 것으로 널리 알려져 있으며(Arvidson and Persson 2000), 아이슬란드 또한 2002년 개정된 납본법에 의거하여 자국 웹 도메인 상의 모든 저작물을 수집하는 정책을 실시하고 있다(Hallgrimsson and Bang 2003; Sigurðsson 2006).

이러한 포괄적 아카이빙의 가장 극단적 예로서, 단일 국가의 경계를 넘어 글로벌 웹 전체를 수집 대상으로 하고 있는 시도가 Internet Archive이다. 단일 국가 영역에 적용되고 있는 포괄적 접근법은 해당 국가의 도메인을 사용하거나, 해당 국가에서나 호스팅되고 있는 사이트들, 그리고 해당 국가와 관련 있는 내용을 담고 있는 사이트들로 수집대상을 제한하고 있는 반면, Internet Archive는 수집 대상의 범위에 어떠한 제한도 두고 있지 않다.

4.3 복합적 아카이빙

마지막으로 여러 가지 웹 아카이빙 방식을 상호보완적으로 운용하는 복합적 접근법이 있다. 앞서 논의하였듯이, 모든 웹 아카이빙 방법들은 제 각기 크고 작은 장점과 단점들을 가지고 있다. 복합적 접근법은 여러 방법론의 혼용을 통해 각각의 단점을 보완하고 장점을 극대화하기 위한 정책이다. 구체적으로는 전체 도메인에 대한 주기적 하베스팅과 선택적 아카이빙이 상호보완적으로 함께 수행되며, 심층 웹의 효율적 수집을 위해 웹 정보자원에 대한 납본이 법제화되는 환경을 지향한다. 노르웨이는 2005년부터 1년에 한 차례씩 자국의 전체 도메인을 대상으로 한 하베스팅을 실시해오고 있다. 하지만 2006년부터는 선택적 수집 방식을 병행하기 시작하여, 매일 단위로 모든 노르웨이 인터넷 신문을, 그리고 주기적으로 노르웨이 인터넷 정기간행물을 수집하고 있다. 이와 함께 이벤트 기반의 수집도 실시되고 있다(Rustad 2005; National Library of Australia. 2008a; 곽승진 외 2008, 71-72). 프랑스 국립도서관 또한 이와 같은 복합적 접근법을 채택하고 있는 대표적인 기관이다. 다음에서는 프랑스 국립도서관의 사례를 통해 복합적 웹 아카이빙 정책에 대해 살펴보고자 한다.

5. 복합적 웹 아카이빙 정책: 프랑스 BnF의 사례

5.1 DADVSI[18)]

프랑스는 2006년 프랑스 국립도서관(Bibliothèque nationale de France, 이하 BnF)과 국립 오디오비쥬얼연구소(National Audiovisual Institute)[19)]의 책임 아래, 납본법 DADVSI에 의거한 법적 납본의 범위를 웹사이트까지로 확대하였다.

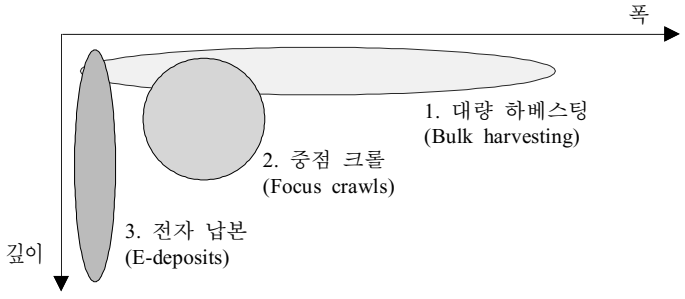
18) Droit d'auteur et droits voisins dans la société de l'information-loi 2006-961.
19) 〈http://www.ina.fr/〉 [citied 2008. 10. 3].

2006년 8월 3일 공포된 납본법 DADVSI는 제39조에서 제47조까지의 타이틀IV에서 웹 정보자원 납본의 법적 근거를 제시하고 있다. 법 제39조에서는 법적 납본의 의무 대상에 "전자적 채널들에 의해 대중들로 커뮤니케이션되는 기호, 이미지, 사운드, 모든 종류의 메시지"를 포함시킴으로써 웹 정보자원을 납본 대상에 포함시키고 있다. 이 법은 최소한 하나 이상의 하이퍼링크로 상호 링크된 웹사이트들의 집합체로 이해되는 "공적 열람 · 검색 영역"을 통해 유통되는 모든 형태의 "온라인 전자 간행물"에 적용된다.

다시 말해 전자적 채널을 이용하여 대중들과 커뮤니케이션할 목적으로 온라인 정보자원을 출판 혹은 생산하는 모든 사람들은 이 법에 따른 납본의 의무를 지게 된다. 하지만 이 법의 적용범위는 프랑스 영토 안에서만 적용된다 - 스위스 국립도서관이 지리적 제약과 상관없이 "스위스에 관한 모든 것"을 수집하는 것과 대조적이다(BnF 2006).

법 41 Ⅱ조에 "권한을 위임받은 기관들은 자동 기술을 사용하거나 또는 생산자와 함께 특정 납본 절차와 계약을 맺음으로써 인터넷으로부터 자원을 수집"할 수 있다고 명시하고 있다. 이에 근거한 BnF는 상호보완적인 대량 하베스팅(bulk havesting), 중점 크롤(focus crawls), 전자 납본(e-deposits)이라는 세 가지 방법을 혼합하여 웹 아카이빙에 나서고 있다. 다음의 〈그림 2〉는 BnF 웹 아카이빙의 범위를 폭(width)과 깊이(depth)로 도식화하고 있다.

이와 같은 BnF 웹 아카이빙 정책은 〈표 2〉로 정리될 수 있다. 법률, 개별 템플릿 및 지침 등 다양한 제도적 근거에 기반 한 여러 가지 수집 방법이 단계별로 적용되고 있음을 볼 수 있다. 다음에서는 각 단계별 세부 절차에 대해 살펴보자.



〈그림 2〉 BnF 아카이브 정책(Illien 2007)

〈표 2〉 BnF 3단계 웹 아카이빙 정책

단계	정책	방법	지침	적용
1단계	일반 정책	대량 하베스트	DADVSI 법	대용량 자동화 및 중점 하베스트 모두에 적용
2단계	중점 정책	중점 하베스트	모든 큐레이터를 위한 일반 지침과 템플릿	모든 수집 단위의 중점 및 이벤트 하베스트에 적용
3단계	주제별 정책	개별 수집	각 단위별 개별 지침	각 단위별 개별 적용

5.2 제1단계: 프랑스 웹사이트들에 대한 대량 자동화 하베스트

대량 하베스트는 로봇에 의해 1년에 1회 수행된다. 2004년 BnF는 Internet Archive와 계약을 맺고 공동으로 프랑스 국가도메인에 대한 연구프로젝트를 착수하였다. 이러한 협력을 통해 BnF는 2004년, 2005년, 2006년 말에 각각 프랑스 도메인 사이트의 스냅샷(snapshot)들을 얻었다. 각 스냅샷은 1억1천8백만에서 1억4천만 개의 파일을 담고 있는데 이는 7테라바이트에 달하는 분량이다. 2007년부터는 BnF 자체로 스냅샷을 수행하고 있으며, 프랑스의 국가 도메인 '.fr'의 등록을 맡고 있는 프랑스 네트워크 정보센터(AFNIC)[20]와 협력하고 있다. 2007년 AFNIC의 리스트를 사용하여 3억 개에 달하는 URL을 목표로 하였고, 약 9테라바이트의 결과물을 얻었다(BnF 2008).

이 단계에서 수집의 범위는 nTLD '.fr'에 포함되는 모든 도메인의 모든 웹사이트와 프랑스 내 다른 콘텐츠를 포괄하며, BnF는 별도의 법적 허가 없이 이러한 수집을 실시할 수 있다. 다만, 수집된 정보자원에 대한 접근은 BnF 내에서만 가능한 것으로 제한되어 있다.[21] 만약 수집된 정보자원 중에 불법·유해정보가 있다 해도 도서관은 이에 대한 책임을 지지 않는다.

또한 BnF는 큰 규모의 웹사이트가 과도하게 수집되는 것을 방지하고자 도메인 당 할당량을 두어, 로봇으로 하여금 동일 도메인에서 1만 개 이상의 URL을 수집하지 않도록 설정하였다. 이는 모든 도메인에게 수집될 동일한 기회를 부여하고자 함이다. 프랑스의 오디오비주얼 유산의 보존을 책임지고 있는 국립오디오비주얼연구소(Institut national de l'audiovisuel, INA)가 오디오비주얼 커뮤니케이션과 관련된 사이트들(대부분 라디오와 TV 사이트들)을 수집하고 BnF가 그 외 다른 모든 사이트들을 책임진다(Cordereix 2008). 현재 BnF는 1단계인 로봇에 의한 대량 자동 하베스트에 가장 우선 순위를 두고 있다.

5.3 제2단계: 중점 하베스트

중점 하베스트는 참고 사서 혹은 주제에 의해 선정된 사이트를 기반으로 한다. 따라서 주제 또는 이벤트 기반의 수집이라고도 한다. 이 단계는 특정 프로젝트 혹은 특정 컬렉션을 대상으로 하는데, 예를 들어, 정부 간행물 사이트, 페스티벌 사이트, 신문 사이트 등이 이에 해당된다. 2006년부터 시작된 BnF의 중점 하베스트는 다양한 주기로 연중 실시된다.

이를 위해 BnF는 사서와 IT전문가 간의 협동작업을 기반으로 하는 디지털 큐레이터 네트워크를 운영하고 있다. 각 수집 단위별로 직원과 리더를 두고, 연중 교육과 워크숍을 실시한다. 단위별 리더들은 1년에 4차례 정기모임을 가지고 운영위원회를 구성한다. 큐레이터들이 이용하는 템플릿의 예는 다음 〈표 3〉과 같다(Illien 2007).

---

20) Association Française pour le Nommage Internet en Coopération. AFNIC는 프랑스 국립 컴퓨터과학·통제연구소(The French National Institute for Research in Computer Science and Control)와 프랑스 정부에 의해 1971년 설립된 비영리 기관으로, '.fr'(프랑스)와 '.re'(레유니온) 도메인의 등록·관리를 책임지고 있다. AFNIC에는 공공 및 민간기관과 개인이 참여하고 있다.

21) 2008년 4월부터는 BnF의 인증된 방문자가 웹 아카이브에 접근할 수 있게 되었다. 이 접근은 국립도서관 내의 연구도서관 열람실 안에서만 가능하다. 이러한 제한은 모든 법적 납본 수집물에 적용된다.

〈표 3〉 템플릿 예제

사이트 유형	크기	변화 · 수명	하베스트 주기	하베스트 범위
정부간행물사이트	대형	안정	수시	심층
페스티벌 사이트	중소형	높은 변화율 · 단기	특정 날짜	심층
신문 사이트	초대형	매우 높은 변화율	주1회: 스냅샷/뉴스	표면
			연1회: 전체 사이트	심층

5.4 제3단계: 납본

1단계와 2단계의 자동 하베스팅이 불가능할 경우, 각 수집 단위들은 각각의 상황에 적합한 수집 정책을 선택하여 수행할 수 있다. 이를 위해서 로봇의 활동을 모니터링하는 담당자는 수집물의 범주와 가치에 대한 명확한 인식을 가지고 있어야 하며, 디지털 큐레이터들의 정책 개발이 요구된다. 이 단계의 대상에는 특수 영역, 주제, 프로젝트 등이 해당되는데, 특정 선거 기간에 이와 관련된 사이트들이 그 예다. 또 다른 경우에는, BnF의 요청에 따라 웹사이트의 생산자가 직접 납본할 수도 있다. DADVSI법은 정보 생산자가 BnF의 하베스팅을 제한하는 로그인, 비밀번호, 혹은 어떠한 형태의 접근 장벽도 사용해서는 안 된다는 점을 명시하고 있다. 또한 법 제50조는 BnF의 수집 요청을 거부한 정보 생산자에게 벌금을 부과할 수 있도록 하고 있다. 하지만 시행은 2009년으로 유예되어 있다.

6. OASIS: 비판적 분석

현재 우리나라의 국가적 수준의 웹 아카이브는 국립중앙도서관의 OASIS 프로젝트에 의해 수행되고 있다. 2004년 첫발을 내딛은 OASIS는 2007년 한 해 동안 10만여 건의 웹 자원을 수집하였으며,[22] 2006년 2월부터 웹사이트(http://www.oasis.go.kr)를 통한 대국민 서비스를 실시하고 있다.

OASIS는 자원수집지침을 통해 수집대상 정보자원의 기준으로 정보자체의 특성과 우리나라와 관련성이라는 두 가지 측면을 제시하고 있다. 첫째, 정보 자체의 특성에 관한 기준으로 "현재 또는 미래의 정보요구에 봉사하는 유용성, 저작자의 평판, 제공된 정보의 유일성, 학술적 내용, 정보의 최신성, 업그레이드 빈도, 접근의 용이성" 등을 제시하고 있으며, 둘째는 우리나라와의 관련성에 관한 기준으로 지리적 경계에 대한 언급 없이 우리나라 사람이 저술한 것인지, 우리나라와 관련된 사회, 정치, 문화, 종교, 과학 또는 경제적으로 중요한 주제인지, 국내외적으로 해당 학문분야에 기여를 한 것인지를 기준으로 제시하고 있다. 이는 수집 대상을 자국 영토 내에 기반 한 웹 정보자원으로 국한하고 있는 프랑스 국립도서관의 정책과 다른 모습이다. OASIS는 수집우선자료와 수집제외자료를 다음 〈표 4〉와 같이 정하고 있다(OASIS 2006a).

22) OASIS 추진경과. [cited 2008.8.3] 〈http://www.oasis.go.kr/intro/intro_history.jsp〉.

〈표 4〉 OASIS 수집 우선자료와 제외자료

수집우선자료	• 중앙정부가 생산한 온라인 디지털자원 • 대학간행물 • 회의자료[23] • 전자저널	• 기증/추천된 온라인 디지털자원 • 최근 이슈가 되는 온라인 디지털자원 • 국내 웹사이트
수집제외자료	• 채팅사이트 • 언론사이트 • 게시판과 뉴스그룹	• 기타 기술적으로 수집이 불가능하며 수집/보존의 가치가 없는 디지털자원

또한 수집 대상을 파일 단위로서 정보 가치를 가지는 개별자원으로서의 "웹문서"와 사이트의 완결된 구조로서 가치를 가지는 "웹 사이트"로 구분하여 구분된 수집과 보존 절차를 적용하고 있다. 수집된 정보자원은 KDC 체계에 따라 분류되고 더블린코어 메타 데이터의 15개 요소를 기준으로 기술된다. OASIS의 이러한 선택적 수집방법은 정보자원의 저작권자와의 사전협의를 통한 계약을 전제로 하고 있다(김유승 2007; OASIS 2006b).

이러한 측면에서 OASIS는 앞서 논의한 '특성별 비교 모델'(그림 1 참조)에서 오른쪽 하단의 ❸영역에 가까운 특징들을 가지고 있다고 할 수 있다. 또한 OASIS를 앞선 사례 유형 분류(표 1 참조)에 대비해 볼 때, 캐나다 국립도서관의 '정적 웹 자원의 선택적 아카이빙'과 네덜란드 국립도서관의 '선택된 상업 정보생산자와의 협동적 계약에 근거한 아카이빙'이라는 두 가지 유형이 복합된 형태의 선택적 아카이빙으로 특징지어 볼 수 있다 – 단 OASIS의 협동적 계약은 비상업적 정보에 집중되어 있다. 이를 종합하면, OASIS의 활동은, BnF의 3단계 웹 아카이빙 정책(표 3 참조)과 비교하였을 때, BnF 웹 아카이빙의 2단계 중점 정책을 소극적으로 적용한 형태로 분류할 수 있을 것이다.

하지만, 앞서 언급하였듯이 웹 정보자원의 납본을 법제화한 캐나다, 그리고 특수한 자국내 출판시장을 전제로 한 네덜란드와 단순 비교하기에는 한계가 있는 것이 사실이다. 다시 말해 웹 아카이빙을 둘러싼 법제도적, 사회적 환경과 출발점이 다르다는 것이다. 더구나, OASIS는 〈그림 1〉 ❸ 영역의 대표적인 예인 미국 의회도서관의 Minerva 프로젝트와는 다른 면모를 지니고 있다. OASIS가 〈표 4〉의 수집지침에 따라 정보자원을 수집한 후 주제별 분류 체계를 적용하는 반면, Minerva는 각각의 독립된 주제별 수집 정책을 채택하고 있다.

이러한 맥락에서, 앞서 웹 아카이빙 비교 분석의 기준으로 삼았던 수집 범위, 방법, 품질의 측면을 살펴볼 때, OASIS는 다음과 같은 문제점들을 지니고 있다.

첫째, OASIS는 수집지침에서 '이슈가 되는 온라인 디지털 자원'을 수집 우선 자료로 하고 있음에도 불구하고, 지난 몇 년간 사회 · 문화

23) 국내외적으로 권위있고 규모가 큰 회의 자료와 정부기관, 전문협회 · 기관, 대학교에서 개최되는 회의자료를 우선적으로 수집한다. 다만 대학학과와 같이 소규모 집단에서 개최되는 회의자료는 수집대상에서 제외된다. 국제회의의 기준은 전체 참가인원 300명 이상, 40% 이상이 외국인 참가자, 5개국 이상의 참가, 3일 이상의 회의기간을 기준으로 한다(이혜원 2004).

적으로 이슈가 되어온 주제별 정보자원들에 대한 수집에 적극적으로 나서고 있지 않다- Minerava의 컬렉션이 2000년 미국 대통령 선거, 2001년 9·11 테러 사건, 2002년 미국 의회 선거 등을 포함하고 있고, 이를 웹 사이트를 통해 이용가능하게 하고 있는 것과 매우 대조적이다.[24] 더구나, 〈표 5〉의 수집물 분류의 예에서 보듯이, 수집된 아이템이 하나도 없는 하위분류가 전체 하위분류의 절반 이상을 차지하고 있어, 웹 정보자원에 대한 KDC 적용의 실효성에 상당한 의문을 남기고 있다. 또한 수집제외 자료에 '기타 기술적으로 수집이 불가능하며 수집/보존의 가치가 없는 디지털자원'을 포함시킴으로서 스스로 다양한 수집 기술의 적용 가능성을 배제시키고 있으며, 앞서 논의한 정보자원의 미래적 가치에 대한 예단이라는 선택적 접근법의 태생적 한계를 그대로 드러내고 있다.

둘째, OASIS는 유일한 국가적 웹 아카이빙 프로젝트라는 위상에도 불구하고 제한적이고 소극적인 수집방법에 머물고 있다. 웹 정보자원에 대한 납본제도가 뒷받침되고 있지 않은 상황에서, 유일무이한 국가대표 웹 아카이빙의 소극적인 수집 정책은 수많은 웹 정보자원의 영구적 손실로 이어질 수 있다. 웹 정보자원의 법적 납본 제도의 도입 시기와 여부가 불확실한 현실을 감안할 때, 자발적 납본 시스템의 시범 실시, 전체 도메인 하베스팅을 위한 타 기관과의 협력적 파일럿 프로젝트 추진 등 다양한 대안을 적극 검토해보아야 할 것이다.

셋째, 수집물의 품질에 관한 문제이다. 〈표 6〉에서 보여주듯이 2008년 10월 현재 OASIS가 이용자들에게 제공하고 있는 수집물 중 웹 문서는 4만4천여 건이 넘는 반면, 웹 사이트는 260여 건에 불과하다. 단일 웹 문서와 사이트의 정보량을 단순 비교할 수는 없음에도 불구하고, 수집 아이템의 수적 측면에서 웹 사이트는 전체 수집 아이템의 약 0.6%에 불과하다. 더구나, 수집된 웹 사이트들의 경우에도 만족스럽지 못한 수평적·수직적 품질을 보이고 있다.

〈표 5〉 OASIS 수집물 분류의 예[25]

총류	웹문서 (284)	총류	243	일반 연속간행물	0
		도서학, 서지학	11	일반학회, 단체, 협회, 기관, 연구기관	1
		문헌정보학	13	신문, 언론, 저널리즘	16
		백과사전	0	일반전집, 총서	0
		강연집, 수필집, 연설문집	0	향토자료	0
	웹사이트 (31)	총류	2	일반 연속간행물	0
		도서학, 서지학	0	일반학회, 단체, 협회, 기관, 연구기관	27
		문헌정보학	2	신문, 언론, 저널리즘	0
		백과사전	0	일반전집, 총서	0
		강연집, 수필집, 연설문집	0	향토자료	0

24) 각주 16) 참조.
25) 출처: OASIS 웹 사이트. [citied 2008.10.7] 〈http://www.oasis.go.kr〉.

〈표 6〉 OASIS에서 제공하고 있는 수집물 수[26)]

	웹문서	웹사이트
총류	284	31
철학	65	3
종교	508	20
사회과학	9093	75
순수과학	695	6
기술과학	31663	79
예술	209	15
언어	403	5
문학	493	11
역사	942	9
합	44355	263

예를 들어 2008년 2월 수집된 '국립중앙도서관' 사이트의 경우(아카이빙 버전 6), 홈페이지 상의 16개 외부 링크가 작동하지 않는 낮은 수평적 품질을 보이고 있다. 반면 2006년 6월 수집된 '한국정보보호산업협회' 사이트는 최상위 네비게이션의 6개 내부 링크가 모두 작동하지 않은 상태로 불완전한 수직적 품질을 나타냈다. 이러한 맥락에서 OASIS는 수집의 범주 및 방법에 대한 연구와 함께, 수집물, 특히 웹 사이트의 품질에 대한 전반적인 점검과 개선이 필요한 것으로 판단된다.

7. 마치며

이상에서 웹 아카이빙을 범위, 방법, 품질의 세 가지 측면으로 비교 분석하고 이를 유형화시키는 논의를 진행하였다. 그리고 이를 바탕으로 최적의 웹 아카이빙 접근법에 대한 논의의 일환으로 BnF의 복합적 아카이빙 정책을 살피고, 국립중앙도서관의 OASIS를 비판적으로 분석해보았다. OASIS는 스스로 밝히듯이 국가대표 웹 아카이브로서 "미래 디지털 세대를 위한 현세대의 디지털 지적 문화유산의 수집/보존"을 목적으로 한다. 하지만, 현재 OASIS의 모습은 이러한 목적 달성을 위해 상당한 개선 노력이 필요한 것으로 판단된다. 이에 본 연구는 결론에 갈음하여 OASIS 프로젝트, 나아가 우리나라 웹 아카이빙 정책의 발전을 위한 두 가지 방안을 제시하고자 한다.

첫째, 복합적 웹 아카이빙 정책의 검토이다. 기존의 선택적 아카이빙 접근법과 함께, Minerva 방식의 주제별 아카이빙, 도메인 중심의 포괄적 아카이빙, 그리고 법적 납본의 전단계로서 자발적 납본제도 등 다양한 웹 아카이빙 접근법의 복합적 운용 정책에 대한 검토가 시급하다. 다시 말해, 국가 대표 아카이브로서 수집 범위를 가능한 가장 넓고 심도 깊게 설정하고, 광

26) 출처: OASIS 웹사이트. [citied 2008. 10. 11] 〈http://www.oasis.go.kr〉.

범위한 수집 범위를 효과적으로 망라할 수 있는 다양한 아카이빙 접근법의 적용이 필요하다는 것이다.

우리가 알고 있는 사실은 누구도 수집하지 않는 엄청난 양의 웹 정보자원이 생산되고 있다는 것과 그것이 소실될 위기에 처했다는 것이다. 이러한 맥락에서 도서관을 비롯한 모든 정보자원을 다루는 기관들은 웹 정보자원의 보존을 위해 그 수집범위와 방법을 다각도로 확장하여야 한다. 장기적인 전망으로, 국가적 웹 아카이빙은 웹 정보자원에 대한 법적 납본 제도를 기반으로 하여야 한다. 21세기 우리 일상의 모든 것이 표출되고 있는 웹 공간에 대한 선택된 정보자원을 대상으로 하는 수집이 아니라 이를 실질적으로 반영해낼 수 있는 납본이 필요하다는 주장은 결코 새로운 것이 아니다. 유네스코는 이미 2000년 개정된 납본지침을 통해 온라인 디지털 정보를 납본 대상에 포함시킬 것을 권고하였고(Larivière 2000), 2003년에는 인터넷 정보자원의 보존 및 이용의 필요성을 명시한 '디지털유산보존헌장(Charter on the Preservation of the Digital Heritage)'을 채택한 바 있다(UNESCO 2003).

국가적 웹 아카이빙의 목적은 '사회, 정치, 문화, 종교, 과학 또는 경제적으로 중요한 주제' 혹은 '최신성, 희소성, 유용성'이 검증된 정보자원의 수집 · 보존에만 있는 것은 아니다. 과거 그리고 오늘의 국가적 웹 공간이 어떠한 모습이었는가를 미래 세대가 보고, 확인하고, 판단할 수 있도록 온전한 근거를 제공하는 것 또한 국가적 웹 아카이빙의 목적이다. 웹 아카이빙에서 수집물은 그 사회를 반영하고 현재적 가치 혹은 대중성과 상관없이 가능한 모든 다양한 문화를 담아낼 수 있어야 한다는 것이다. 도서관 서가에 대철학자의 책과 통속 소설이 함께 자리하고 있듯 웹 정보자원의 수집에도 같은 철학이 적용되어야 한다. 선택적 수집은 맥락 없는 단편적 표본들의 모음으로써 법적 납본의 대안이 될 수 없다. 무엇이 가치 있는 것인가를 결정하는 일은 다음 세대의 몫이다.

대량 하베스트 방식은 이러한 철학을 웹 공간으로까지 확장해 적용시킬 수 있는 기술적 해법을 제시하고 있다. 하지만 웹 정보자원의 태생적 특성으로 인해 대량 하베스트 방식은 또 다른 측면의 한계들을 드러내고 있다. 따라서 여러 가지 웹 아카이빙 접근법의 한계를 상호 보완할 수 있는 복합적 아카이빙 정책이 대안으로서의 가치를 가진다 할 수 있고, 이에 BnF의 웹 아카이빙 정책은 시사하는 바가 크다.

둘째, 타 기관과의 협력적 웹 아카이빙 전략의 수립이다. 웹 아카이빙은 실로 단일 기관이 감당하기에 벅찬 일이다. 웹 정보자원의 다원성과 다양성, 크기와 변화라는 태생적 특성이 이를 더욱 어렵게 만들고 있다. 따라서 다양한 기관들의 협력 체계 구성이 필요하다는 주장은 일찍이 제기되어 왔다(김유승 2007; 서혜란 2004, 17). 앞서 살펴보았듯이, 호주의 PANDORA는 10여 기관의 협력으로 운영되며, Screen Sound Australia는 음악, 필름 관련 사이트를, Australia War Memorial은 호주 군사역사 관련 사이트를 수집, 보존하는 책임을 지고 있다. 프랑스의 경우에도 BnF와 국립오디오비주얼연구소가 분야별로 웹 아카이빙의 책임을 나누고 있다. 국립중앙도서관의 웹 아카이빙도 타 기관들과의 협력에 의해 수행되어야 한다—이러한 분업과 협력의 형태는 지역별로 또는 주제별로 구성될 수 있을 것이다.

또한 국립중앙도서관을 중심 허브로 하는 개별 기관별 협력일 수도, 공공 및 민간 영역의 다수 주체가 동등한 자격으로 참여하는 콘소시엄의 형태가 될 수도 있을 것이다.

지금 이 순간도 웹은 진화하고 있으며, 그 속의 정보자원들도 빠르게 변화·발전하고 있다. 적극적인 아카이빙의 노력 없이는 오늘 웹 공간에 펼쳐지는 우리들 삶의 기록들과 정보자원이 언제 기억의 저편으로 사라질지 누구도 장담하지 못한다. 미래를 내다보는 웹 아카이빙 정책의 수립은 미래 세대를 위해 오늘을 사는 우리가 짊어져야 할 책임이자 약속이어야 한다. 웹 아카이빙 정책을 둘러싼 이론적 논의와 실천이 시급한 이유는 바로 이 때문이다.

참 고 문 헌

곽승진, 최재황, 조영주, 류희경. 2008. 디지털 납본에 대한 보상 체계 요구. 한국도서관·정보학회지』, 39(2): 65-83.

김유승. 2007. 웹 아카이빙의 법·제도적 문제에 대한 고찰.『한국문헌정보학회지』, 41(3): 5-24.

서혜란. 2004. 웹 아카이빙의 성과와 과제.『한국비블리아』, 15(1): 5-22.

송병호. 2005. 진본성 확보를 위한 전자기록물 관리 방안.『한국비블리아학회지』, 16(2): 43-59.

이소연. 2004. 디지털유산의 장기적 보존: 국가정책수립을 위한 제언.『기록학연구』, 10: 27-64.

이혜원. 2004. 온라인 디지털 자원 구축 사례: 국립중앙도서관을 중심으로.『2004년 디지털 유선보존에 관한 기초연구 보고서』. 147-160.

임진희. 2006. 전자기록의 장기보존을 위한 보존정보패키지(AIP) 구성과 구조.『기록학연구』, 13: 41-90.

Arvidson, Allan and Persson, Krister. 2000. The Kulturarw3 Project－The Royal Swedish Web Archiw3e－An example of "complete" collection of web pages. *66th IFLA Council and General Conference*, Jerusalem, Israel. [citied 2008. 10. 4] 〈http://www.ifla.org/IV/ifla66/papers/154-157e.htm〉.

Baeza-Yates, R. A. and C. Castillo. 2004. Crawling the infinite Web: Five levels are enough. In Workshop on algorithms and models for the Web-Graph WAW 2004. Rome, Italy: Springer Verlag.

Bergmark, D. 2002. *Collection syntesis*. Paper presented at the 2nd ACM/IEEE-CS Joint Conference on DIgital libraries, Portland, July 14-18.

Bergmark, D., C. Lagoze, and A. Sbityakov. 2002. *Focused crawls, tunneling, and digital libraries*Paper presented at the 6th European Conference on Research

and Advanced Technology for Digital Libraries, Rome, Italy, September 16-18.

BnF. 2006. Web archiving at BnF. [citied 2008. 10. 3] 〈http://www.bnf.fr/PAGES/version_anglaise/depotleg/pdf/BnFnews200609.pdf〉.

BnF. 2008. Legal deposit:five questions about Web Archiving at BnF. [citied 2008. 10. 3] 〈http://www.bnf.fr/PAGES/version_anglaise/depotleg/dl-internet_quest_eng.htm〉.

Cordereix, Pascal. 2008. The legal deposit of Audiovisual and Multimedia material in France: the example of the Audiovisual Department of the National Library of France(Bibliothèque national de France/BnF). *World Library and Information Congress: 74th IFLA General Conference and Council*. Québec, Canada.

Cruse, P., Eckman, C., and Kunze, J. 2003. *Web-based government information: Evaluating solutions for capture, curation, and preservation*. Okland: California Digital Library.

Day, Michael. 2003. Collecting and preserving the World Wide Web: A feasibility study undertaken for the JISC and Wellcome Trust. UKOLN, University of Bath.

Electronic Publication Pilot Project team and Electronic Collections Committee. 1996. Electronic Publication Pilot Project (EPPP) Final Report. [cited 2008. 10. 1] 〈http://dsp-psd.pwgsc.gc.ca/Collection/SN3-331-1997E.pdf〉.

Hallgrimsson, Þorsteinn and Bang, Sverre. 2003. Nordic Web Archive. *3rd ECDL Workshop on Web Archives*. Trondheim, Norway. [citied 2008. 10. 3] 〈http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=hallgrimsson〉.

Henriksen, Birgit. 2001. Legal Deposit from the Internet in Denmark : Experiences with the Law from 1997 and the Need for Adjustments. *Preserving the Present for the Future: Strategies for the Internet conference*. The Royal Library, Copenhagen 18th -19th of June 2001. [cited 2007. 7. 25] 〈http://www.deflink.dk/upload/doc_filer/doc_alle/1023_BNH.doc〉.

Illien, Gildas. 2007. Re-Inventing Collection Development Policy in the Age of Web Archiving: Experience of the National Library of France. BnF. [cited 2008. 8. 2] 〈http://www.ku.edu.tr/ku/images/LIBER/LIBER_ILLIEN_2008.ppt〉.

Johnson, David., David, Post. 1996. Law and borders: The rise of law in cyberspace. *Stanford Law Review*, 48(1367).

Kim, Heejung and Hyewoon, Lee. 2007. Development of Metadata Elements for Intensive Web Archiving. 『정보관리학회지』, 24(2): 143-160.

Koehler, W. 2004. "A longitudinal study of

Web pages continued: a consideration of document persistence." *Information Research*, 9(2): 174.

Koerbin, Paul. 2004. PANDORA: Australia's web archive. Library Science Talks 2004. Swiss National Library/CERN. [citied 2008. 10. 2] 〈http://libraryscience talks04.web.cern.ch/libraryscience talks04/pandora.ppt〉.

Larivière, Jules. 2000. *Guidelines for Legal Deposit Legislation*. Paris: UNESCO.

Lasfargues, France., Oury, Clément., and Wendland, Bert. 2008. Legal deposit of the French Web: harvesting strategies for a national domain. *IWAW '08*. Aarhus, Denmark. [citied 2008.10.3] 〈http://iwaw.net/08/IWAW2008-Lasfargues.pdf〉.

Lawrence, S., D. M. Pennock, G. W. Flake, R. Krovertz, F. M. Coetzee, E. Glover, F. Å. Nielsen, A. Kruger, and C. L. Giles. 2001. "Persistence of Web references in scientific research." *Computer*, 34(2): 26-31.

Lyle, J. A. 2004. Sampling the Umich.edu domain. *The 4th International Web Archiving Workshop*. Bath, UK.

Masanès, Julien. 2002. Archiving the deep Web. *2nd International Workshop on Web Archives*. Rome, Italy. [citied 2008. 10. 4] 〈http://bibnum.bnf.fr/ecdl/2002/BnF/BnF.html〉.

Masanès, Julien. 2005a. *Collecting the hidden Web*. In J. Masanés (Ed.), Web Archiving. New York: Springer Verlag.

Masanès, Julien. 2005b. "Web Archiving Methods and Approaches: A Comprartive Study." *Library Trends*, 54(1): 72-90.

National Library of Australia. 2008a. Legal deposit. [cited 2008. 7. 30] 〈http://www.nla.gov.au/padi/topics/67.html〉.

National Library of Australia. 2008b. Web archiving [cited 2008. 10. 1] 〈http://www.nla.gov.au/padi/topics/92.html〉.

National Library of the Netherlands. 2004. The archiving system for electronic publications: The e-Depot. [citied 2008. 10. 3] 〈http://www.kb.nl/kb/dnp/e-depot/dm/dm-en.html〉.

OASIS. 2006a. OASIS 소개: 온라인디지털자원 선정 지침. [cited 2008. 8. 3] 〈http://www.oasis.go.kr〉.

OASIS. 2006b. OASIS 시스템: 수집보존프로세스. [cited 2008. 10. 8] 〈http://www.oasis.go.kr〉.

Phillips, Margaret. 2003. Collecting Australian Online Publications. PANDORA. [cited 2008. 9. 4] 〈http://pandora.nla.gov.au/bsc49.doc〉.

Phillips, Margaret. 2005. "What should we preserve? The question for heritage libraries in a digital world." *Library Trends*, 54(1): 57-71.

Rustad, Kjersti. 2005. Our digital heritage as source material to end-users: collection of and access to net publications in

The National Library of Norway. *World Library and Information Congress: 71th IFLA General Conference and Council* Oslo, Norway. [citied 2008. 10. 3] 〈http://www.ifla.org/IV/ifla71/papers/151e-Rustad.pdf〉.

Sigurðsson, Kristinn. 2006. Archiving the Icelandic Web. *ICA/SUV* Seminar. Reykjavík, Iceland. [citied 2008.10.3] 〈http://www2.hi.is/solofile/1009995〉.

UNESCO. 2003. Charter on the Preservation of the Digital Heritage. The 32nd session of the General Conference of UNESCO, 17 October 2003.