정보관리연구, vol.42, no.3 2011, pp.95-116

웹 아카이브 OASIS의 현황에 관한 연구

A Research on the OASIS, an Web Archive in Korea, Revisited

윤 정 옥*
Cheong-Ok Yoon

차 례

1. 머리말
2. OASIS의 '자원 수집 지침'과 수집 자원의 현황
3. 주제별 분석
4. 최신 자료와 많이 본 자료
5. 맺음말
· 참고문헌

초 록

이 연구에서는 국립중앙도서관이 구축·운영하는 웹 아카이브인 OASIS 콘텐츠의 특성과 문제점을 살펴보았다. 2011년 5월 20일부터 6월 10일 사이 OASIS 웹사이트를 통해 제공되는 주제별 디렉토리에서 철학, 종교 등 5개 주제와 사회과학과 기술과학 소주제의 수집 웹사이트 55종 및 '최신 자료'와 '많이 본 자료' 리스트를 분석하였다. OASIS는 수집 콘텐츠의 주제 분포의 편중, 웹사이트 저작자/발행자 권위 및 학술적 가치의 근거 미약, 아카이빙 대상의 정의 부적합, 수집 및 이용 통계 처리의 모호성 등 문제점이 나타났고, 이에 대한 포괄적이며 체계적인 개선책 마련이 시급한 상황이다.

키 워 드

오아시스, 웹 아카이브, 국립중앙도서관, 디지털자원

* 청주대학교 문헌정보학과 교수
(Professor, Dept. of Library and Information Science, Cheongju University, jade@cju.ac.kr)
• 논문접수일자: 2011년 6월 9일
• 최종심사(수정)일자: 2011년 7월 18일
• 게재확정일자: 2011년 7월 20일

ABSTRACT

The purpose of this research is to examine the characteristics and problems of the OASIS, an web archive, developed and operated by the National Library of Korea. An analysis of 55 websites collected in seven subject areas, including 'Philosophy', 'Religion', 'Literature', 'Social Sciences-Politics', and 'Technology-Technology' in the OASIS website shows the following problems: lack of balance in subject distribution, lack of authority or scholarly value of some contents, unclear application of selection criteria for personal creators/publishers, ambiguous definition of 'archiving', and inappropriate counting of collection and usage statistics. Some comprehensive and systematic efforts to improve the OASIS contents seem to be urgently needed.

KEYWORDS

OASIS, Web Archive, National Library of Korea, Digital Resources

1. 머리말

1.1 연구의 필요성과 목적

우리나라 국립중앙도서관의 OASIS(Online Archiving & Searching Internet Sources)는 "국가 지식자원의 수집과 보존 책임"을 가진 국가대표도서관이 "가치 있는 인터넷 자료를 국가적인 차원에서 수집·축적하여 미래 세대에 연구 자료로 제공과 온라인 디지털자원 수집 보존의 표준모델 연구 등에 그 목적을 두고" 추진한 프로젝트이다(국립중앙도서관, OASIS 2009a).

국립중앙도서관은 2001년 '온라인 저작물 수집보존 시범 시스템'을 구축하고, 2004년부터 웹 자료를 수집하기 시작하여 2006년 2월 OASIS 홈페이지에서 대국민 서비스를 시작하였다. 2004년 OASIS 프로젝트를 시작할 당시 국가적 웹 아카이브의 필요성이나 웹 아카이빙의 정책적 목표나 전략 개발 등에 대한 연구들이 수행되었으나(서혜란 2004; 최원태 2005; 이혜원 2004), 실제로 OASIS가 서비스를 시작한 이후 국립중앙도서관이 수집하고 OASIS 홈페이지를 통해 제공하는 웹 자료 콘텐츠의 내용과 품질이 과연 "국가 지식자원"으로서 적정한 수준을 달성하고 있는지 체계적으로 평가한 연구는 별로 없었다.

그동안 미국의 RLG와 국립기록관리관(NARA)이 제시한 디지털 아카이브의 평가도구인 "TRAC(Trustworthy Repositories Audit & Certification)"을 사용하여 OASIS를 검토한 연구(이소연 2008)와 OASIS의 선택적 웹 자원 수집 방법에 대한 비판적 분석 연구

(김유승 2007; 2008)가 있었다. 최근 이용자의 관점에서 OASIS의 '문학'과 '사회과학' 분야 콘텐츠와 서비스를 검토한 연구(윤정옥 2010)는 OASIS 콘텐츠가 주제 분포의 편중, 저작자/발행자 편향성과 권위의 근거 미약, 정보의 유일성과 최신성 결여, 웹 문서와 웹사이트의 중복 수집, 학술적 가치의 선정 근거 결여 등 문제점을 가지고 있음을 지적한 바 있다.

최근 국립중앙도서관은 2010년부터 OASIS를 위해 "국제적으로 통용되고 있는 수집로봇인 Heritrix를 국내 온라인 자료의 성격에 맞게 구축하여 기존에는 수집하지 못하던 자료들까지 수집"하고, "수집된 온라인 자료를 기술하는 메타데이터 포맷도 DC(Dublin Core)에서 보다 확장성이 뛰어난 MODS(Metadata Object Description Schema)로 변경·적용하여 이용자가 사용하기에 편리한 환경을 구현"하였다고 보고하였다. 또한 2010년도에는 웹문서 14만4,452건, 웹사이트 7,365건을 수집하여 총 15만1,817건의 공개용 온라인 자료를 수집하였다고 하였다(국립중앙도서관 2011, 66-68). 그러나 이와 같은 수집로봇의 도입으로 인한 수집방법의 변화, MODS라는 새로운 자료 기술 방법의 적용 및 수집 웹 자료의 증가 등이 과연 "가치 있는" 웹 자료의 확보, 질적 가치 향상 및 서비스 개선으로까지 이어졌는지는 알기 어렵다.

이러한 맥락에서 이 연구는 OASIS가 제공하는 웹 자료 콘텐츠의 현황을 분석하고, 운영주체인 국립중앙도서관이 보다 체계적인 평가나 분석을 시행하는 데 도움이 될 수 있도록 이용자의 관점에서 개선점을 제안하는 것을 그 목적으로 하였다.

1.2 연구의 내용과 방법

이 연구에서는 2011년 5월 20일부터 6월 10일까지의 3주 동안 OASIS 홈페이지에서 제공되는 웹 자료, 즉 수집된 웹사이트 350종 가운데 55종을 조사하여 콘텐츠 및 접근성과 관련된 특성을 다음과 같이 검토하였다:

첫째, 콘텐츠와 관련하여 '한국십진분류법(KDC)'으로 분류된 주제별 디렉토리 리스트에서 웹사이트의 간략정보 및 상세정보에서 최초 및 최근 아카이빙 일자 등을 확인하여 주제별로 〈표〉를 구성하였고, 각 웹사이트의 아카이빙 버전을 일일이 확인하는 과정을 거쳤다. 또한 리스트에서 링크가 활성화되어 있는 경우에는 현행 운영되는 웹사이트로도 직접 들어가 정보의 최신성, 저작자의 평판과 같은 특성을 살펴보았다.

둘째, 접근성과 관련하여 OASIS에서 제공하는 사이트 검색 기능, '최신자료' 및 '많이 보는 자료' 리스트의 기능 등을 살펴보았다.

2. OASIS의 '자원 수집 지침'과 수집 자원의 현황

2.1 OASIS의 '자원 수집 지침'

OASIS의 '자원 수집 지침'은 현재 또는 미래의 정보요구에 대한 유용성, 저작자의 평판, 제공된 정보의 유일성, 학술적 내용, 정보의 최신성, 업그레이드 빈도, 접근의 용이성을 일차적인 수집 기준으로 삼고 있다. 이 기준에는 수집되는 웹 자료가 한국과 관련되었고, 사회, 정치, 문화, 종교, 과학, 경제적 중요성을 가지며, 한국의 저자가 작성하고, 권위 있는 한국 사람이 국내외적으로 해당 학문 분야에 기여할 수 있어야 한다는 조건을 포함하고 있다. 이러한 '자원 수집 지침'에 부합하는 수집 우선 자료로는 중앙정부가 생산한 온라인 디지털자원, 대학간행물, 회의자료, 전자저널, 기증/추천된 온라인 디지털자원, 최근 이슈가 되는 온라인 디지털자원 및 국내 웹사이트를 포함하며, 수집 제외 자료는 채팅사이트, 언론사이트, 게시판과 뉴스그룹, 그리고 기타 기술적으로 수집이 불가능하며 수집/보존의 가치가 없는 디지털자원을 포함하고 있다(국립중앙도서관, OASIS. 2009d).

국립중앙도서관은 『2008 국립중앙도서관연보』에서 OASIS를 위해 그동안 "연구보고서, 간행물, 정책자료, 통계자료 등 주로 웹사이트 내 공개자료실에서 제공되는 웹문서와 신행정수도, 선거사이트 등 모든 주제의 웹사이트"를 수집하였다고 보고한 바 있다(국립중앙도서관 2009). 이후 『2010 국립중앙도서관연보』에서는 "특히 지역축제, 국회의원선거, 행정도시(세종시) 건설 등 한국 내 사회적 이슈가 되는 디지털자원도 수집"하였고, 대표적인 예로는 "제18대 국회의원 후보자 개인홈페이지, 지방자치선거, 신행정수도건설위원회 홈페이지 등을 수집"했다고 보고하였다(국립중앙도서관 2011, 66-68). 이로써 OASIS에서는 기본적인 '자원 수집 지침'의 기준에 맞을 뿐 아니라 매년 시사적인 의미로 관심 대상이 되는 다양한 주제의 웹 자료들을 수집한다는 원칙을 알 수 있다.

2.2 OASIS의 수집 자원의 현황

2.2.1 수집 실적

〈표 1〉은 국립중앙도서관이 2004년 처음 웹 자료(웹사이트와 웹문서)를 수집하기 시작한 이래로 2010년까지의 매년 수집 실적을 보여주고 있다(국립중앙도서관 2011, 68). OASIS는 2004년에 처음 4만96건의 웹 자료를 수집한 이래 매년 꾸준히 수집 자원의 수가 증대함

〈표 1〉 연도별 웹 자료 수집 실적

구분	2004	2005	2006	2007	2008	2009	2010	합계
웹 자료	40,096	50,259	63,725	100,512	112,070	66,086	151,817	584,565

을 볼 수 있다. 2008년에는 11만2,070건이었던 수집 실적이 2009년에는 6만6,086건으로 급격히 감소하였지만, 다시 2010년에 15만1,817건으로 증가하였다. 2010년 말 누적 웹 자료의 총수는 모두 58만4,565건에 이르고 있다. 이 통계에서 웹사이트와 웹 문서의 건수가 별도로 제시되지는 않았다.

그러나 〈표 1〉의 수집 실적은 크게 두 가지 의문점을 갖게 한다.

첫째, 수집한 58만4,565건이 모두 고유한 웹 자료인가 하는 점이다. 뒤에서 논의할 것처럼 분명 2009년부터 2010년 사이에 각 주제별 웹사이트의 수가 통계상으로는 증가되었음에도 실제 각 주제 리스트로 들어가 보면 웹사이트의 최초 아카이빙 일자, 즉 최초의 수집일자가 2010년인 웹사이트의 수가 증가된 수치와 일치하지 않는다. 말하자면 실제 웹사이트의 증가가 아니라, 이미 수집해 놓은 웹사이트를 추가 아카이빙 함으로써 호주의 PANDORA에서 '아카이브된 인스턴스'로 정의된 웹사이트의 특정 시점 '스냅샷'의 증가가 아닌가 하는 점이다. '아카이브된 인스턴스(archived instance)'란 아카이브에 추가된 타이틀의 "단일 스냅샷(a single snapshot)"이나 카피를 말한다. 예를 들어, 연속간행물 타이틀이 새로운 이슈를 추가할 때처럼, 많은 타이틀이 변화하는 콘텐츠를 캡처하기 위하여 한번 이상 아카이브에 카피된다. 이럴 때 각각의 새로운 카피는 '아카이브된 인스턴스'라고 불린다(National Library of Canada 2009c).

둘째, OASIS에서 2011년 현재 접근 가능한 웹 자료의 수는 전체 보유 자원 58만4,565건의 5.9%인 350건밖에 되지 않는다. 2011년 5월 현재 국립중앙도서관이 OASIS에서 공개하고 있는 온라인자료는 〈표 2〉에 보는 바와 같이 350건의 웹사이트뿐이다. 이전에 OASIS 관련

〈표 2〉 OASIS 수집 자원의 현황(인용 2011. 5. 25)

주제	2009년 10월 9일			2011년 5월 25일		2009-2011
	웹 문서	웹사이트		웹사이트		증가율
총류	416	31	12.0%	40	11.4%	29.0%
철학	44	3	1.2%	5	1.4%	66.7%
종교	467	20	7.7%	18	5.1%	-10.0%
사회과학	10,828	78	30.1%	111	31.7%	42.3%
순수과학	705	6	2.3%	7	2.0%	16.7%
기술과학	32,933	80	30.9%	104	29.7%	30.0%
예술	1,760	15	5.8%	25	7.1%	66.7%
언어	521	5	1.9%	7	2.0%	40.0%
문학	636	11	4.2%	14	4.0%	27.3%
역사	1,131	10	3.9%	19	5.4%	90.0%
합계	49,441건	259종	100.0%	350종	100.0%	35.1%

※ 윤정옥. 2010. 12. 〈표 3〉에서 '주제별 디렉토리' 데이터를 근거로 구성함.

연구가 수행되었던 2009년 9월 당시에는 4만 9,076건의 웹 문서와 259종의 웹사이트를 공개하고 있었으나, 현재는 웹 문서는 공개하지 않고 있다.

〈표 2〉는 2011년 5월 25일 현재 OASIS의 주제별 디렉토리에 공개된 웹사이트의 통계를 지난 2009년 10월 9일 당시 공개되었던 웹 문서 및 웹사이트 통계와 비교한 것이다. 2009년에는 OASIS에서 49,441건의 웹 문서와 259종의 웹사이트를 구별하여 공개하였으나, 2011년에는 웹 문서는 공개하지 않고 있다. 2011년 현재 공개된 웹사이트는 모두 350종으로 2009년보다 91종이 증가하여, 35.1%의 증가율을 보였다. 분명 공개 종수는 증가한 것으로 나타나고 있다.

2.2.2 수집 자원의 주제별 분포

OASIS의 주제별 디렉토리는 웹 문서와 웹사이트를 '한국십진분류법(KDC)'에 따라 분류하고 있다. 2009년 주제별 분포를 보면 259종의 웹사이트 가운데 '기술과학' 주제 웹사이트가 80종(30.9%), '사회과학' 주제 웹사이트가 78종(30.1%)으로 가장 높은 비중을 차지하고 있다. 2011년에는 350종의 웹사이트 가운데 '사회과학' 주제 웹사이트가 111종(31.7%), '기술과학' 주제 웹사이트가 104종(29.7%)으로 순위가 바뀌었다. 이 두 주제에 속하는 웹사이트들이 2009년에는 전체의 61.0%, 2011년에는 61.4%를 각각 차지하고 있어, 주제 편중이 아직도 매우 심하다고 할 수 있다.

2009년 당시 '철학' 주제 웹사이트는 3종(1.2%), '언어'는 5종(1.9%), '순수과학'은 6종(2.3%) 등은 매우 낮은 비중을 차지하였고, 2011년에도 크게 달라지지 않았다. '철학' 주제 웹사이트는 5종(1.4%)으로, '언어' 주제 웹사이트는 7종(2.0%)으로 약간 증가하였으나, '순수과학'은 7종으로 늘었으나 전체 비중은 2.0%로 2009년보다 다소 낮아졌다.

앞서 국립중앙도서관이 2010년에 7,365건의 웹사이트를 수집하였다고 하는데 현재 OASIS 웹사이트에 350건을 어떤 기준으로 선별하여 공개하고 있는지 알기 어렵다. 현재까지 수집된 전체 웹사이트의 주제별 분포도 이 공개된 웹사이트들에 반영된 것과 같은 주제의 편중 현상을 보이고 있는지 분석이 필요하다.

2.2.3 수집 자원의 주제별 증가율

주제별 디렉토리에 공개된 웹사이트의 종수는 2009년에서 2011년 사이에 35.1% 증가하였다. 그 가운데 웹사이트 증가율이 가장 높은 주제는 '철학'과 '예술' 두 분야이다. '철학' 주제 웹사이트는 2009년 3종에서 2011년 5종으로 늘어남으로써 명목증가율은 66.7%에 달하였다. 그러나 워낙 수집 종수가 미미하여 전체 주제별 분포에는 기여하는 바가 크지 않다. '예술' 주제 웹사이트는 2009년 15종(5.8%)에서 2011년 25종(7.1%)으로 증가하여 증가율은 '철학'과 마찬가지로 66.7%이지만, 전체 주제별 분포는 1.3% 늘어났다.

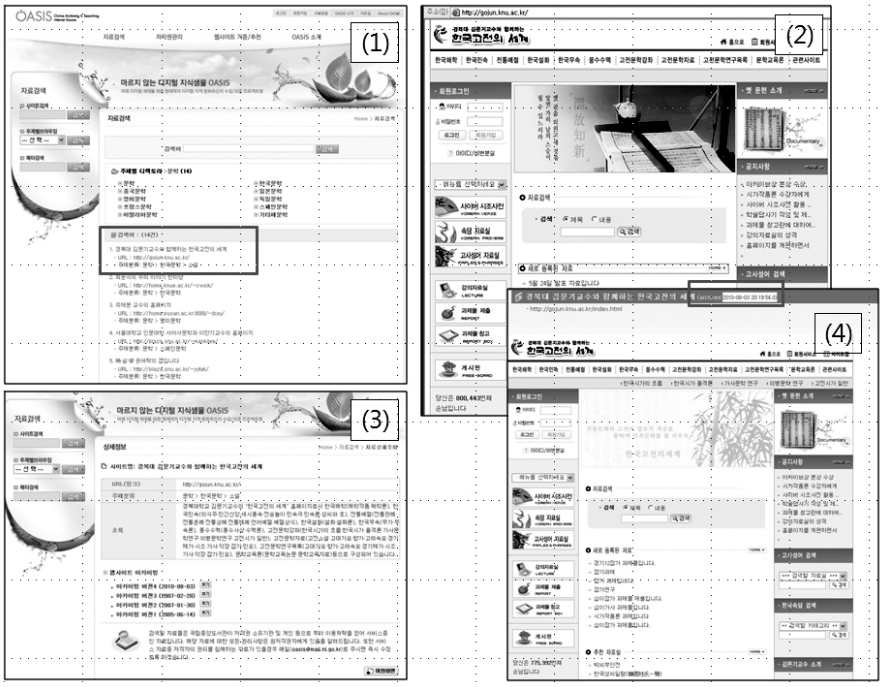
한편 '종교' 주제 웹사이트는 2009년 20종(7.7%)이 수집되었으나, 2010년에는 18종

(5.1%)만이 수집되어 10.0%의 감소율을 보였고, 전체 주제별 분포는 2.5% 줄어들었다. 다른 아홉 개 주제 분야의 웹사이트들이 조금씩이라도 수집 종수가 늘어난 데 비하여 '종교' 부문만 감소한 것으로 나타났다.

'사회과학'과 '기술과학'은 주제별 분포 비중이 높기도 하지만, 2009년과 2011년 사이 증가율 또한 높다. '사회과학'은 33종의 웹사이트가 추가되어 2009년 대비 42.3%의 증가율을 보였고, '기술과학'은 24종의 웹사이트가 추가되어 30.0%의 증가율을 보였다. 이 두 분야는 기존 웹사이트 수가 절대적으로 많고, 증가율도 높아 여전히 전체 주제별 분포에서 압도적인 위치를 차지하고 있다.

2.3 주제별 디렉토리와 서지정보

〈그림 1〉의 (1)은 2011년 5월 25일 주제별 디렉토리의 '문학' 분야에 아카이빙된 웹사이트 리스트이다. 이 리스트에서 각각의 사이트명, URL, 주제분류 항목을 볼 수 있고, 여기에서 URL은 활성화되어 있어 (2)와 같이 해당 사이트로 바로 들어갈 수 있다. 그런데 (1)에서 보는 웹사이트 리스트는 정렬의 기준을 알 수 없다. 다음에서 살펴보겠지만, 각 주제 리스트에서 웹사이트들은 사이트명의 가나다순으로 배열된 것도 아니고, 최초 아카이빙 일자순으로 배열된 것도 아니다. 또한 이용자가 웹사이트들을 재정렬 할 수 있도록 허용하지도 않는다.



〈그림 1〉 '문학' 주제 웹사이트 리스트와 상세정보(인용 2011. 5. 25)

〈그림 1〉의 (3)는 첫 번째 사이트인 '경북대 김문기교수와 함께하는 한국고전의 세계'의 '상세정보'로 앞의 리스트에서 사이트명을 클릭하면 볼 수 있다. '상세정보'는 사이트명과 더불어 URL(링크), 주제분류, 초록 등을 포함하고 있는데, 여기에서 URL은 활성화되어 있지 않다. 국립중앙도서관은 2010년부터 OASIS의 메타데이터를 더블린코어에서 MODS로 변환하여 기술하고 있다고 하였는데, 2010년 10월 당시에는 '상세정보'에서 메타데이터 요소로 URL, 수집파일명, 저작자, 주제분류, 요약, 발행자, 생성일 및 UCI 등 항목을 포함하고 있었던 것에 비하여, 현재는 단 네 개의 항목만 포함하고 있다. 상세정보 항목 축소 외에는 MODS 변환 이후 어떤 점이 달라졌는지 알 수 없다.

한편 (4)는 '상세정보'의 '웹사이트 아카이빙' 리스트에서 '아카이빙 버전 4'의 '보기'를 클릭한 결과로서, 2010-08-03 20:19:54.00에 아카이빙된 이 웹사이트의 스냅샷을 볼 수 있다.

3. 주제별 분석

여기에서는 주제별 디렉토리에 제시된 350종의 웹사이트 가운데 55종을 선정하여 2011년 5월 25일부터 6월 7일 사이에 주제별로 검토하였다. 웹사이트의 선정 기준은 KDC의 10개 주제 분야에서 철학, 순수과학, 언어와 같이 수집 종수가 10개 미만인 주제는 모두 살펴보고, 사회과학, 기술과학과 같이 수집 종수가 많은 웹사이트는 OASIS 선정 기준에서 언급한 내용을 직접 관찰할 수 있는 소주제의 웹사이트들을 임의로 선정하여 살펴보았다.

각 주제 리스트에서 웹사이트의 간략정보를 확인하고, 상세정보에서 최초 및 최근 아카이빙 일자 등을 확인하여 〈표 3〉을 구성하였으며, 각 웹사이트의 아카이빙 버전을 일일이 들어가서 확인하는 과정을 거쳤다. 또한 리스트에서 링크가 활성화되어 있는 경우에는 웹사이트로도 직접 들어가서 확인하였다.

3.1 철학

'철학' 주제 웹사이트는 2009년 단 3종이 수집되었으나, 2011년 현재는 5종으로 증가하였다. 그러나 실제로 어떤 웹사이트가 새로 추가된 것인지는 알 수 없다.

여기 선정된 웹사이트들 가운데 '대한철학회'는 상세정보에 아카이빙 일자 및 아카이빙 버전이 포함되어 있지 않다. '대한철학회'의 웹사이트로 직접 들어가 확인한 결과로는 이 학회가 "한국칸트학회라는 이름으로 1964년 『哲學硏究』 창간호를 발행하면서 출발"하였고, 현재 한국의 철학회지들 중에서 가장 오래된 역사를 기록하는 『哲學硏究』(학술진흥재단 등재지)를 간행하고 있었다. 따라서 '대한철학회'의 웹사이트는 권위 있는 학회의 공식 웹사이트로서 철학이라는 주제 분야의 권위를 보여준다고 할 수 있으나, 그 밖의 사이트들에 대해서는 그렇다고 하기 어렵다.

〈표 3〉 OASIS 웹사이트의 현황(55종)(인용 2011. 5. 25 – 6. 7)

주제 구분	웹사이트명	아카이빙		
		최초	최근	총 회수
철학 (5종)	광운대학교 산업심리학과 유태용교수 홈페이지	2005.8.29	-	1
	김성환교수와 함께 하는 동양철학의 세계 …	2005.8.29	-	1
	덕성여대 심리학과 김정호교수 홈페이지	2005.8.29	1	1
	한국기독교상담심리치료학회	2006.2.9	2007.2.16	2
	대한철학회	-	-	-
종교 (10종)	세계선교연구원	2005.11.28	2007.2.22	2
	박노권교수의 함께 만들어가는 세상	2005.8.24	-	1
	이승구교수의 개혁신학 이야기	2005.8.24	-	1
	정종훈교수의 생활신앙, 생활윤리	2005.8.30	-	1
	카톨릭 교리마당(CateForum)	2005.8.29	2010.7.25	4
	국제기독교언어문화연구원	2005.11.11	2010.5.7	3
	한국기독교연구소	2006.2.24	2007.10.4	2
	기독교학문연구회	2005.11.22	2010.1.14	3
	민중신학연구소	2005.11.21	-	1
	한국기독교장로회신학연구소	2005.11.22	2007.2.23	2
사회과학-정치학 (8종)	전라남도 목포시 가선거구 여인두후보 홈페이지(2010년 6·2 지방선거)	2010.6.30	-	1
	김세균 홈페이지: 이론과 정치	2005.8.25	2010.7.9	4
	경상남도 이달곤후보 홈페이지(2010년 6·2 지방선거)	2010.6.3	-	1
	이동복(전 자민련 국회의원)	2005.6.20	2007.6.12	2
	한국여성정치연구소	2007.3.20	-	1
	전가림교수의 강의도움 사이트	2005.8.30	2007.6.17	2
	강성구 개인 홈페이지(제17대 국회의원선거)	2004.4.10	-	1
	한국여성유권자연맹	2006.10.19	-	1
순수과학 (7종)	인간유전체기능연구사업단	2005.8.23	2007.6.13	2
	고석구교수의 홈페이지	2005.8.24	2007.6.8	2
	버섯하고 놀자	2005.7.8	2007.6.13	2
	한국전기화학회	-	-	-
	한양대학교 양자광기능물성연구센터	2005.11.11	2010.4.17	3
	한국천문학회	-	-	-
	한국생물정보학회	2006.4.13	2008.3.4	2
기술과학-기술과학 (4종)	간문자교수의 홈페이지	2005.8.24	-	1
	한국생산기술연구원	2005.5.4	2010.5.28	2
	한국표준과학연구원	2005.5.3	2010.8.3	4
	한국과학기술출판협회	2006.9.11	-	1
문학 (14종)	경북대 김문기교수와 함께하는 한국고전의 세계	2005.6.14	2010.8.3	4
	최운식의 우리 이야기 한마당	2005.8.30	-	1
	유제분교수의 홈페이지	2005.8.29	-	1
	서울대학교 인문대학 서어서문학과 이만기교수의 홈페이지	2005.8.30	-	1

주제 구분	웹사이트명	아카이빙		
		최초	최근	총 회수
문학 (14종)	時·삶·學 윤여탁의 집입니다	2005.8.30	-	1
	황국명교수의 문학세계	2004.12.1	2005.8.26	2
	최희섭교수의 홈페이지	2005.8.26	-	1
	글쟁이 권소연	2005.7.21	2010.4.3	3
	한국현대문학회	2005.9.1	2007.6.13	2
	광마클럽-MaKwangsoo	2005.7.5	2007.6.13	2
	한국수필창작문예원 에세이코리아	2004.12.6	2007.1.25	2
	시와 반시	2004.11.23	2007.3.22	2
	백담사 만해마을	2007.9.14	-	1
	시인세계	2004.12.7	2010.3.30	2
언어 (7종)	양병곤교수의 홈페이지	2005.8.29	2010.4.3	3
	윤웅진교수의 홈페이지	2005.8.29	-	1
	서울대학교 국어교육연구소	2006.2.16	2007.2.20	2
	한글학회	2004.11.18	2010.5.7	2
	이야기 한자여행	2005.7.7	2010.2.12	3
	국립국어원	-	-	-
	팬코리아영어교육학회	2004.11.15	2007. 6.12	3

2005년 8월 29일 수집되어 아카이브된 '광운대학교 경영대학 산업심리학과 유태용교수 홈페이지'는 그의 전공분야가 '인사선발, 선발검사 개발, 직무분석, 산업교육, 직무수행평가, 효용성 분석' 임을 밝히고 있으며, 이 웹사이트는 '응용심리학이론'으로 분류되고 있다. '덕성여대 심리학화 김정호교수 홈페이지'는 저자가 '명상, 스트레스, 복지, 건강 등의 주제'를 다루고 있으며, '심리학'으로 분류되고 있다. 이 두 웹사이트 및 '김성환교수와 함께 하는 동양철학의 세계' 웹사이트는 현재는 운영되지 않는다. 이들은 모두 개인의 웹사이트로서 강의자료 등을 주로 수록하고 있는데, 모두 2005년 8월 29일 단 한 차례씩만 아카이빙 되었다. 이들이 개인 저작자로서 각기 해당 분야의 주제적 권위를 얼마나 가지고 있는지는 확인할 수 없다.

'철학' 주제의 웹사이트 5종 가운데 '대한철학회'를 제외한 4종의 웹사이트는 모두 2005년에서 2006년 사이 처음 아카이빙 되었고, 이들 가운데 2009년부터 2011년 사이에 추가된 것으로 확인할 수 있는 웹사이트는 없다.

3.2 종교

'종교' 주제 웹사이트는 모두 18종이 공개되고 있다. 전체 웹사이트 수가 많지도 않은 동시에 이들은 모두 '기독교'로 분류되어 있어, 주제의 편중성을 보여주는 대표적인 분야이다. 이 '종교' 주제 웹사이트들 가운데 리스

트의 앞부분에 위치한 10종의 웹사이트를 살펴보았다. 이들 가운데 '세계선교원'은 장로회신학대학 소속기관이며, '국제기독교언어문화연구원'은 남서울대학 소속기관이다. 한편 목원대학교 박노권교수, 국제신학대학대학원 이승구교수, 연세대학교 교목인 정창권교수의 개인 홈페이지들이 수집되어 있으며, 이들은 모두 단 1회씩만 아카이빙 되어 있다. '카톨릭교리마당'은 김웅태신부가 운영하는 웹사이트로서 이례적으로 4회나 아카이빙이 되어 있는데, 각 아카이빙 버전의 수록내용은 방명록, 공지사항, 게시판의 내용 외에는 거의 변동이 없다.

앞서 '철학' 주제와 마찬가지로 이들이 개인이나 기관 저작자로서 어떤 주제적 권위 혹은 학술적 가치를 갖는 콘텐츠를 제공하는 지 확인할 근거가 없다.

3.3 사회과학

'사회과학' 주제에 수집된 웹사이트는 2009년 78종에서 2011년 현재 111종으로 증가하였다. 〈표 4〉에 보는 바와 같이 수집된 웹사이트들은 사회과학 8종, 경제학 25종, 정치학 8종, 법학 8종, 풍속, 민속학 21종, 사회학, 사회문제 20종, 행정학 7종, 교육학 10종 및 국방, 군사학 4종이다. 통계학 주제 웹사이트는 하나도 수집되지 않았다. 이 주제의 모든 웹사이트를 전수조사하기 어려우므로 앞서 OASIS가 2010년 지방선거, 행정수도 등 관련 웹자료를 수집하였다고 보고한 것에 근거하여 '사회과학-정치학' 소주제를 살펴보았다.

앞서 〈표 3〉에서 살펴본 '사회과학-정치학' 주제로 수집된 웹사이트는 모두 8종이다. 이들 가운데 2010년 6월 2일 지방선거와 관련이 있는 것은 '전라남도 목포시 가선거구 여인두후보 홈페이지(2010년 6·2 지방선거)-다음블로그'와 '경상남도 이달곤후보 홈페이지(2010년 6·2 지방선거)' 단 2종의 웹사이트이다. 또한 이 8종의 웹사이트들 가운데 국회의원 관련 웹사이트로는 전자민련 국회의원인 '이동복 홈페이지'와 제17대 국회의원 '강성구 개인 홈페이지'만이 포함되어 있다. 여기 수집된 웹사이트들이 2010년 6·2 지방선거에서 혹은 국회의원들 가운데서 얼마나 중요한 정보적 가치와 대표성을 갖는지 알기 어렵다.

〈표 4〉 '사회과학' 분야 웹사이트의 소주제 분포(111종) (인용 2011. 6. 7)

소주제	수집 종수	소주제	수집 종수
사회과학	8	통계학	0
경제학	25	사회학, 사회문제	20
정치학	8	행정학	7
법학	8	교육학	10
풍속, 민속학	21	국방, 군사학	4

한편 국립중앙도서관이 수집하였다고 한 '신행정수도건설위원회' 홈페이지는 OASIS 웹사이트에서 검색되지 않으며, 만약에 이 웹사이트가 수집되었다면 포함될 만한 '사회과학' 분야 주제 브라우징 리스트 어디에도 들어있지 않다.

미국의회도서관 웹 아카이브(The Library of Congress Web Archives. Minerva)는 "지정된 주제에 대한 웹 기반 정보를 대표하기 위하여 주제전문가가 선정한 아카이브 된 웹사이트의 컬렉션"으로서 구축됨에 따라, 선정된 주제에 대해서는 매우 포괄적인 웹 자료를 수집하고 있다. 예를 들어 2002년부터 매 2년 간격으로 선거 관련 웹 아카이브 컬렉션을 구축하고 있으며, 그 가운데 'The United States Election 2008 Web Archive'는 미국의 대통령, 의회 및 주지사 선거에 관련된 2,200여 사이트를 선별해 놓은 컬렉션으로서, 2007년 4월부터 2008년 12월까지 수집된 것들이다. 이 컬렉션은 대통령, 의회, 주지사 후보자, 주정부, 정당, 선거감시자, 공공정책 및 정치 주창 그룹, 후보자와 관련 선거 블로그, 그리고 2008년 선거와 관련된 그 밖의 "잡다한(Miscellaneous)" 웹사이트들까지 수집해 놓고 있다(Library of Congress Web Archives 2011a).

우리나라의 OASIS는 미국의회도서관 웹 아카이브와 같이 "다음 세대의 연구자를 위해서 디지털 자원을 평가, 선정, 수집, 목록, 접근 제공 및 보존"하겠다는 중차대한 사명(Library of Congress Web Archives 2011b)을 천명하지는 않았다. 그러나 2010년 6·2 지방선거와 관련하여 단지 2종의 웹사이트를 수집해 놓고 이를 "대표적인 예"로서 공공연히 공표하는 것은 그리 적절하지는 않다.

한편 수집된 웹사이트들 가운데 '전가림교수의 강의도움 사이트'는 2회 아카이빙 되었으나 '아카이빙 버전2(2007.6.17)'은 '아카이빙 버전1(2005.8.30)'과 동일한 시점의 화면을 보여주며 글자가 깨진 상태로 아카이빙 되어 있다.

3.4 순수과학

'순수과학' 분야의 웹사이트는 모두 7종이 수집되어 있다. 이들 가운데 '한국전기화학회'와 '한국천문학회'는 리스트에서 사이트명을 클릭하면 '본 자료는 저작권자의 요청으로 국립중앙도서관 안에서만 볼 수 있습니다'는 메시지가 나오므로, 다른 사이트들과 같이 상세정보로 들어갈 수 없고, 아카이빙 현황을 알 수 없다. 그러나 리스트에서 URL을 클릭하면 바로 해당 사이트로 들어갈 수가 있다. 따라서 위의 메시지의 '본 자료'란 웹사이트가 아니라, 웹사이트의 '상세정보'를 의미하는 것으로 이해할 수밖에 없다. 왜냐면 실제로 저작권자가 웹사이트 자체는 누구나 볼 수 있도록 웹상에 공개해 놓고 있으면서, 웹사이트의 아카이브된 정보를 국립중앙도서관 안에서만 보도록 요청하였다는 것은 이해하기 어렵기 때문이다. 그렇다 하더라도, 저작자가 겨우 단지 사이트명,

활성화되지 않은 URL, 주제분류, 초록 및 웹사이트 아카이빙 일자만을 제공하는 '상세정보' 접근을 국립중앙도서관 안으로 제한하도록 요구했다는 것 또한 설명이 되지 않는다.

'순수과학'은 2009년 10월 당시 6종이 수집되어 있었으나, 2011년 5월 현재는 7종으로 증가하였다. 그러나 〈표 3〉의 현황을 보면 '한국전기화학회'와 '한국천문학회'는 '상세정보'에 대한 접근이 제한됨에 따라 최초 및 최근 아카이빙 일자를 알 수 없으나, 나머지 5종의 웹사이트 모두 2005년부터 2006년 사이 수집되었고, 2009년과 2010년 사이 수집된 것은 없다. 또한 최신 아카이빙도 2007년과 2008년 사이 이루어졌다. 다만 2005년 처음 수집되었던 '한양대학교 양자광기능물성연구센터'만 유일하게 2010년 4월 17일 다시 한 번 아카이빙이 이루어졌을 뿐이다.

이처럼 웹사이트 최초 아카이빙과 최신 아카이빙 일자를 확인한 결과는 역시 '순수과학' 분야에서도 2009년과 2011년 사이 웹사이트 수 증가를 설명하지 못한다.

3.5 기술과학

'기술과학' 분야에는 모두 104종의 웹사이트가 수집되어 있다. 〈표 5〉에 보는 바와 같이 소주제로 기술과학 4종, 농업, 의학 10종, 건축공학 7종, 전기공학, 전자공학 11종, 제조업 3종, 의학 16종, 공학, 공업이론 43종, 기계공학 7종, 화학공학 4종, 가정학 및 가정생활 2종이 수집되었다. 이들 가운데 '기술과학-기술과학' 분야에 수집된 4종의 웹사이트를 살펴보았다. 수집된 웹사이트에는 한국생산기술연구원 등 세 곳의 기관 홈페이지와 더불어 호남대학교 의상디자인학과 '간문자교수의 홈페이지'가 포함되어 있다. 한국생산기술연구원과 한국표준연구원의 웹사이트는 2005년에 처음 아카이빙 되었으나 2010년 한 차례씩 더 아카이빙 되었다.

'기술과학-기술과학' 주제는 단지 4종의 웹사이트를 갖고 저작자의 주제적 권위 등을 논하기는 어렵다. 그러나 수집 웹사이트 종수가 적은 만큼 수집된 특정 웹사이트의 가치에 대하여 이용자가 동의할 수 있는 어떤 근거가 제시되면 좋을 듯하다. 이를테면 이들이 비록 로

〈표 5〉 '기술과학' 분야 웹사이트의 소주제 분포(104종) (인용 2011. 6. 7)

소주제	수집 종수	소주제	수집 종수
기술과학	4	의학	16
농업, 농학	10	공학, 공업이론	43
건축공학	7	기계공학	7
전기공학, 전자공학	11	화학공학	4
제조업	3	가정학 및 가정생활	2

봇에 의한 기계적 수집의 결과라 하더라도 메타데이터 요소에 '수집 근거' 등의 항목을 추가함으로써 국가에서 보존할 디지털 자원으로서의 가치를 확인할 수 있게 하는 것이 바람직할 것이다.

3.6 문학

'문학' 분야에서 2011년 5월 25일 현재 공개된 웹사이트는 모두 14종으로 〈표 2〉에서 보는 2009년 10월 현재 수집된 웹사이트 11종보다 3종이 증가한 것이다. 그러나 실제로 이들의 리스트에서 최초 아카이빙 날짜를 보면 14종의 웹사이트 가운데 2004년에 처음 수집된 웹사이트가 4종, 2005년에 수집된 웹사이트가 9종, 그리고 2007년 수집된 웹사이트가 1종이다. 즉 2009년부터 2011년 사이에 새로 수집된 웹사이트는 없다. 그럼에도 불구하고 3종의 웹사이트가 증가한 것으로 나타난 것은 그 근거를 알 수 없다. 다만 이들 가운데 '경북대 김문기교수와 함께하는 한국고전의 세계', '글쟁이 권소연', 및 '시인세계'라는 세 개 웹사이트의 최근 아카이빙 일자가 2010년이므로 이것을 통계에 넣은 것이 아닌가 추정될 뿐이다.

3.7 언어

'언어' 주제로는 모두 7종의 웹사이트가 수집되어 있다. 이들 가운데 '국립국어원' 사이트는 앞서 '한국천문학회' 등과 같이 접근이 제한되어 있어 아카이빙 일자를 확인할 수 없었다. '윤웅진교수의 홈페이지'는 리스트의 링크를 통하여 접근하니 현재 운영되지 않고 있으나, '상세정보'에서는 2005년 8월 29일 현재 아카이빙된 웹사이트에 접근할 수 있었다.

한편 '한글학회' 웹사이트는 '상세정보'에 두 차례의 아카이빙 보기가 제시되고 있으나 이를 각각 '아카이빙 버전1(2004-11-18)', '아카이빙 버전3(2010-05-07)'으로 표시되어 있다. 실제 아카이빙은 3회 이루어졌으나 여기에서 두 개의 버전만 보여주는 것이든지, '아카이빙 버전2'가 '아카이빙 버전3'으로 잘못 표기된 것일 수 있으나 확인이 어렵다.

'이야기 한자여행'은 세 차례의 아카이빙이 이루어진 것으로 되어 있으나, '아카이빙 버전3(2010-02-12)'은 '웹 페이지를 찾을 수 없습니다'는 메시지 화면만 아카이빙 되어있고, '아카이빙 버전2(2007-06-13)'는 '아카이빙 버전1(2005-07-07)'과 동일한 시점의 화면이 아카이빙된 것으로서 일부 문자는 깨진 상태로 나타나고 있다. 즉 아카이빙은 2010년 2월 12일까지 모두 3회가 이루어졌다 해도, 실제 아카이빙된 자료는 2005년 7월 7일 자료 상태에 멈추어 있다. 이 웹사이트는 현재도 운영이 되고 있으며 리스트의 URL로부터 접근이 가능하다. 이 경우는 아카이빙을 동일 웹사이트를 다른 시점에서 새로 접근하여 갱신하는 것이 아니라, 복사하여 화면만 재생한 것이 아닌가 하는 의문이 든다.

한편 여기 수집된 '팬코리아영어교육학회'

는 '영남영어교육학회'의 웹사이트로서 이 학회가 우리나라 언어 분야의 가장 대표적인 학회인지 알 수 없다.

'언어' 분야에서도 주제별 브라우징 리스트에서 확인이 어려운 '국립국어원'을 제외한 6종의 웹사이트 모두 2004년부터 2006년 사이 최초 아카이빙된 것으로 2010년 이후 새로 수집된 웹사이트는 보이지 않는다.

4. 최신 자료와 많이 본 자료

OASIS 홈페이지 초기화면에서는 '최신 자료' 및 '많이 본 자료'의 리스트를 보여주며, 이들은 각각 100종의 웹사이트의 사이트명을 제공하고 있다. 여기에서는 이들 가운데 상위 15종의 웹사이트들의 상세정보를 각각 확인하여, 최근 아카이빙 일자, 조회 수 등을 살펴보았다.

4.1 최신 자료

〈표 6〉은 OASIS 홈페이지에서 2011년 5월 25일 현재 '최신 자료'의 최상위 15종을 보여준다. 그러나 여기에서 '최신 자료'의 '최신성' 기준은 무엇인지 알 수 없다. 웹사이트의 최초 아카이빙 일자도 기준이 아니고, 최근 아카이빙 일자도 기준이 아니다. 이 웹사이트 15종의 최근 아카이빙 일자는 〈표 6〉의 오른쪽 컬럼에서 보는 것처럼 모두 2010년 7월에서 10월 사이이지만, 날짜순 정렬은 아니다.

〈표 6〉 OASIS 최신 자료(인용 2011. 5. 25)

	사이트명	최근 아카이빙
1	문학사랑	2010.10.6
2	ChunWoo=월간 문학세계, 계간 시세계, 도서출판 천우	2010.10.6
3	양자 광기능 물성 연구센터	2010.10.9
4	대한건축학회 2003 학생작품전 = AIK 2003 Student Architectural Exhibition	2010.10.6
5	유네스코와유산 = UNESCO & HERITAGES	2010.10.14
6	김세균 홈페이지: 이론과 정치	2010.7.9
7	우수저류 활용연구센터	2010.7.8
8	언론중재위원회	2010.7.16
9	환경정의	2010.7.12
10	한밭교육박물관	2010.7.9
11	한국항공운항학회(KSAF)	2010.7.29
12	한국표준과학연구원	2010.8.3
13	대한건축학회 2003 학생작품전	2010.10.6
14	한국천문학회	2010.7.28
15	불교환경연대	2010.7.26

'최신자료' 리스트의 첫 번째 웹사이트인 '문학사랑'과 두 번째 웹사이트인 'ChunWoo=월간 문학세계, 계간 시세계, 도서출판 천우'의 최근 아카이빙 일자는 2010년 10월 6일이고, 세 번째 웹사이트인 '양자 광기능 물성 연구센터'의 최근 아카이빙 일자는 2010년 10월 9일이다. 또한 다섯 번째 웹사이트인 '유네스코와 유산'의 아카이빙 일자는 2010년 10월 14일이다. 말하자면 최신 아카이빙 일자가 순위 정렬의 기준은 아님을 알 수 있다.

한편 '최신자료' 리스트에서도, 각 사이트의 '상세정보'에서도 '이전화면'으로 이동할 수 없다. 따라서 '최신자료' 리스트에서 어느 사이트의 '상세정보'를 보고 난 후에는 다시 OASIS 초기화면으로 돌아가서 또다시 '최신자료' 리스트로 가야 한다.

4.2 많이 본 자료

OASIS에서는 '많이 본 자료' 또한 100건의 사이트명과 조회 수를 리스트로 제공하고 있다. 〈표 7〉은 2011년 5월 25일 이들 가운데 최상위 15종의 사이트명을 각각 최근 아카이빙 일자와 조회 수를 확인하여 작성하고, 이후 6월 6일 다시 조회 수를 확인한 것이다. 그러나 '많이 본 자료'의 기준은 의문스러운 점이 있다.

첫째, '많이 본 자료'의 기간이 설정되어 있지 않다. 말하자면 하루, 일주일, 혹은 한 달 동안 등과 같이 일정한 기간 내 조회 수를 명시하지 않아, 언제부터 언제까지 기간 동안 '많이 본 자료'를 제공하는 것인지 명확하지 않다.

둘째, 조회 수의 계수가 정확하지 않았다. 이 연구에서 처음 '많이 본 자료' 리스트의 웹

〈표 7〉 많이 본 자료(인용 2011. 5. 25, 2011. 6. 6)

	사이트명	최근 아카이빙	5.25조회	6.6조회
1	문학사랑	2010.10.6	30	31
2	ChunWoo=월간 문학세계, 계간 시세계, 도서출판 천우	2010.10.6	24	19
3	국립중앙도서관	2008.2.21	12	14
4	한국인구학회	2008.12.19	12	14
5	양자 광기능 물성 연구센터	2010.10.9	9	4
6	유네스코와유산 = UNESCO & HERITAGES	2010.10.14	9	8
7	김세균 홈페이지: 이론과 정치	2010.7.9	8	10
8	영암왕인문화축제	2008.4.2	8	5
9	Lee Won Taek's Home Page	2007.6.12	8	5
10	대한민국국방부	2009.5.1	7	8
11	별난물건박물관	2005.11.21	6	7
12	경북대 김문기교수와... 한국고전의 세계	2010.8.3	6	5
13	대한건축학회 2003 학생작품전	2010.10.6	6	5
14	국립광주박물관	2010.7.9	6	5
15	한국교과서연구재단	2010.7.8	5	5

사이트명과 최근 아카이빙 일자를 확인하기 위하여 1회 클릭하면 조회 수가 2회씩 증대하여, 연구자가 들여다보고 있는 사이에 실시간으로 계속 사이트 순위가 바뀌었다. 1위인 '문학사랑'을 처음 보았을 때는 조회 수가 24회였으나 이 연구를 위해 검토하는 사이에 30회가 되었다. 분명 많이 본 자료 순위가 실시간으로 추적되는 것은 사실이지만, 조회 수 계수 방법에는 문제가 있는 듯하였다.

한편 '많이 본 자료'의 조회 수를 확인하기 위하여 2011년 6월 6일 다시 리스트를 검토한 결과는 모두 조회 수가 변동되었고, 최상위 1-4위를 제외하고는 순위도 변동되었다. 2011년 5월 25일 당시 조회 수 30회로 1위였던 '문학사랑'은 6월 6일 현재 조회 수 31회로 여전히 1위이다. 그러나 2011년 5월 25일 당시 조회 수 24회로 2위였던 'ChunWoo'는 19회로 조회 수가 줄었고, 5위였던 '양자 광기능 물성 연구센터'도 조회 수가 9회에서 4회로 줄었다. 이처럼 특정 사이트의 조회 수가 5월보다 6월에 줄어든 것을 보면 아마도 월 단위로 조회 수를 계수하여 '많이 본 자료'를 공개하는 것으로 보인다.

'많이 본 자료'에서 열거된 100종의 웹사이트는 OASIS에서 공개하는 350종의 웹사이트의 28%에 해당한다. 2011년 6월 6일 현재 이들 중 상위 5-6위 정도의 웹사이트만 10회 이상 조회 수를 기록하였으며, 나머지 대다수의 웹사이트들은 조회 수가 매우 적다. 리스트 상의 레코드 번호 12-21의 10종의 웹사이트는 5회씩 조회되었고, 레코드 번호 22-52의 31종의 웹사이트는 4회씩, 레코드 번호 53-75의 24종의 웹사이트는 3회씩 조회되었다. 레코드 번호 76부터 100까지는 단 2회씩 조회되었을 뿐이다. 만약 OASIS가 한 달 단위로 조회 수 순위를 매긴다면 5월 최고 빈도 조회를 기록한 '문학사랑'의 조회 수를 6월 첫째 주 현재 이미 넘어섰으므로 조회 수가 더 높아질 것이라고 기대할 수 있으나, 향후 확인이 필요하다.

5. 맺음말

OASIS는 국립중앙도서관이 지난 2004년부터 구축하여 2006년부터 일반에게 공개하고 있는 국가 디지털 자원 아카이브이다. 2011년 5월 20일부터 6월 10일 사이 이 연구에서 살펴본 OASIS는 그동안 수집된 콘텐츠에 관련된 몇 가지 문제점이 있는 것으로 나타났다.

첫째, 현재 OASIS 웹사이트에 공개된 웹 자료의 수량이 적을 뿐만 아니라 공개 자료의 선정기준이 알려져 있지 않다. 앞서 국립중앙도서관이 2010년도에 웹문서 14만4,452건, 웹사이트 7,365건을 수집하여 총 15만1,817건의 공개용 온라인 자료를 수집하였다고 했음에도 불구하고, 2011년 6월 현재는 웹사이트 350종만을 공개하고 있다. 이것은 2010년 수집자료 총건수의 0.2%, 수집 웹사이트 총수의 4.7%에 해당하는 자료들이며, 〈표 1〉에 따른 OASIS 누적 수집자료 총 건수 58만4,565건의 0.06%에 지나지

않는다. 물론 수집된 모든 웹자료를 공개할 필요는 없다. 그러나 그렇게 극히 적은 수량만을 일반 이용자들 대상으로 공개했을 경우에는 이들의 공개 기준이 무엇인지 명시해 줄 필요가 있다.

둘째, OASIS 웹사이트 공개 자료의 형태가 제한되어 있다. 〈표 2〉에 보는 것처럼 2010년 10월 당시 OASIS에서는 웹사이트뿐만 아니라 웹 문서도 공개하고 있었다. 당시에도 공개기준은 알 수 없었고, 웹 문서가 웹사이트의 내용과 상당 부분 중복된다는 지적이 있었다(윤정옥 2010). 그러나 공개된 자료 형태로는 개별 웹 문서 49,441건에 대한 접근을 허용하였으며, 웹 문서들 가운데 '많이 본 자료' 및 '최신 자료'의 리스트도 제공하고 있었다. 그러나 2011년 6월 현재 웹 문서는 더 이상 공개하지 않고, 웹사이트만 공개하고 있다. 앞서와 같은 이유로 웹 문서를 공개대상에서 제외한 이유가 무엇인지 이용자들에게 공지해 줄 필요가 있다.

셋째, OASIS 수집 통계의 정확성을 확인하기 어렵다. OASIS는 매년 증가된 웹자료의 통계에서 전적으로 새로운 웹 자료가 추가되었는지, 단지 기존 웹 자료의 아카이브된 인스턴스만 추가된 것인지 구별하기 어렵다. 따라서 앞에서 살펴본 주제별 디렉토리의 여러 주제 분야에서 웹사이트 수는 증가하였으나, 그 가운데 어느 것이 새로 추가된 것인지 알 수 없다는 문제가 있었다.

앞에서 지적한 바와 같이 호주의 국가 웹 아카이브 PANDORA 웹사이트는 아카이브 규모와 매월 증가 통계를 제공하는데, 매월 새롭게 추가된 타이틀 수와 기존에 수집된 웹 자료의 새로운 스냅샷으로서 인스턴스 수를 구별하고 있다. 2011년 5월 26일 현재 통계는 2011년 5월 한 달 동안 아카이브된 타이틀 수는 216종 추가되었고, 아카이브된 인스턴스 수는 733건이 증가하였음을 알 수 있다(National Library of Australia 2011a).

OASIS도 호주의 PANDORA와 같이 실제 새로 추가된 웹자료와 인스턴스를 구별함으로써 정확한 웹 자료의 증가 통계를 제공할 필요가 있다. 만약에 OASIS가 2010년에 수집했다는 웹사이트 7,365종이 모두 새로운 것이었다면, 한 달 평균 이삼백여 타이틀을 새로 수집하여 아카이브하고 있는 PANDORA보다도 세 배가량 많은 웹사이트들을 연간 수집한다는 것이니 그 증가율이 대단하다고 하겠다. 그러나 그 증가율이 단순한 인스턴스 추가를 의미한다면 의미는 크게 달라질 것이다.

넷째, OASIS 웹사이트는 이용 통계를 공개하지 않고, 이용 빈도와 관련된 '많이 본 자료' 리스트의 기간을 설정하지 않고 있다. OASIS는 일일방문자수 및 누적 방문자수를 공개하지 않는 한편 호주의 PANDORA는 매우 상세한 이용 통계를 보여준다(National Library of Australia 2011b). 월별, 일별 및 일일 시간대별로 방문 수, 방문 페이지 수, 방문 히트 수 및 대역너비(bandwidth) 통계를 보여줄 뿐만 아니라, 방문자의 국가별 통계, 호스트 IP 통계, 로봇/스파이더 방문자, URL-페이지, 운영

체계(OS), 브라우저 통계 등 다양한 방문자 분석 통계를 보여준다.

국립중앙도서관의 OASIS 웹사이트가 진정으로 "가치 있는 인터넷 자료를 국가적인 차원에서 수집·축적하여 미래 세대에 연구 자료로 제공"하는 정보원의 역할을 하고 있는지(국립중앙도서관, OASIS 2009a), 이용통계를 통해서 확인할 필요가 있다. 만약에 "가치 있는" 자료를 제공함에도 불구하고 이용자들이 별로 없다면, 그 원인을 분석하고 방문을 유인할 방안을 모색해야 할 것이다. 그러나 제공하는 자료자체의 "가치"가 결여되어 있고, 이용자도 별로 없다면, 보다 근본적인 자료 수집 정책을 점검해 보아야 할 것이다.

다섯째, OASIS가 제공하는 웹자료, 즉 수집된 디지털 콘텐츠 "가치"의 적절성을 의심할 만한 여러 가지 근거가 있다. 이 연구에서는 주제별 디렉토리 분류에 따라 '철학' 5종, '종교' 10종, '사회과학-정치학' 8종, '순수과학' 7종, '기술과학-기술과학' 4종, '문학' 14종 및 '언어' 7종의 웹사이트를 포함한 55종의 웹사이트를 살펴보았다. 이들은 공개된 350종의 웹사이트의 육분의 일에 해당하며, '사회과학-정치학'과 '기술과학-기술과학'을 제외한 나머지 다섯 개 주제는 선정된 웹사이트 전체를 살펴본 것이다. 이 웹사이트들은 다음과 같은 특성을 가진다:

(1) OASIS에 수집된 웹사이트의 주제 분포가 매우 편중되어 있다. 우선 주제 분포의 편중성은 전체 350종의 웹사이트 가운데 '사회과학' 분야가 111종으로 31.7%를 차지하고, '기술과학' 분야가 104종으로 29.7%를 차지하여 두 주제 분야만으로 61.4%가 구성되고, 나머지 일곱 개 주제 분야의 웹사이트들이 38.6%를 차지하고 있다는 데서 나타난다. 또한 어떤 주제 안에서는 선정된 웹사이트의 주제들이 편중되어, 예를 들어 '종교' 분야의 웹사이트 10종은 모두 기독교 자료이며, 불교, 천도교 등 다른 주제의 웹사이트는 수집하지 않았다.

(2) 수집된 웹사이트의 저작자/발행자 권위 및 학술적 가치의 근거가 미약하다. 여기에서 분석된 55종의 웹사이트 가운데 '철학' 분야 3종, '종교' 분야 3종, '사회과학-사회과학' 분야 6종, '순수과학' 분야 1종, '기술과학-기술과학' 분야 1종, '문학' 분야 9종, 그리고 '언어' 분야 3종을 포함하는 26개 웹사이트는 개인이 운영하는 것으로 이들은 대부분 교수이고, 일부 정치인, 종교인도 포함되어 있다. 그러나 앞에서 각 주제 분야를 분석한 데서도 언급한 바와 같이 몇몇 교수의 강의 자료나 신상잡기가 들어있는 홈페이지, 한 두명 국회의원의 정치활동 보고 홈페이지 등 여기 포함된 웹사이트들의 '저작자의 평판'이나 이들이 제공하는 '정보의 학술적 내용'이 과연 국가적 디지털 자산으로서의 가치를 가지는 것인지 근거를 확인하기 어렵다.

(3) 이 연구에서 검토한 55종의 웹사이트 가운데 최초 아카이빙 일자가 2010년인 것은

'사회과학-정치학' 분야의 2010년 6.2 지방선거 관련 웹사이트 단 2종뿐이다. 나머지는 대부분 2004년부터 2006년 사이에 최초로 아카이빙 된 웹사이트들이다. 이들 가운데 일부는 2010년에 다시 아카이빙 된 경우도 있지만, 대부분은 최초 아카이빙 버전만 그대로 공개되고 있다. 이 연구에서 검토하지 않은 나머지 웹사이트들에 최신 수집 자료가 얼마나 포함되어 있는지는 알 수 없다.

(4) 전체적으로 OASIS 홈페이지에서 이용할 수 있는 자료의 수와 정보의 범위는 감소하였다. 웹 문서를 이제는 공개하지 않음에 따라 접근 가능한 웹 자료의 수가 대폭 감소되었을 뿐 아니라, 공개된 각 웹 자료의 '상세정보' 제공 항목 수도 감소하였다. 이전에는 더블린코어 메타데이터에 기반하여 URL, 수집파일명, 저작자, 주제분류, 요약, 발행자, 생성일 및 UCI 등의 요소를 사용하고 있었으나, 지금은 MODS로 변경하였다고 하는데, 제공하는 요소는 사이트명, URL, 주제분류, 초록 및 아카이빙 일자의 4개 항목으로 줄어들었고, 이용 면에서 무엇이 개선된 점인지 알 수 없다.

지금까지 살펴본 OASIS 공개 웹 자료의 현황을 보면 OASIS가 "마르지 않는 디지털 지식 샘물"로서, 그리고 "미래 디지털 세대를 위한 현세대의 디지털 지적 문화유산의 수집/보존 프로젝트"로서 제 역할을 과연 하고 있는 것인지 의문을 갖게 된다. 2004년에 OASIS 프로젝트를 시작하여 2011년 현재에 이르기까지 매년 몇 만에서 십여 만 건까지의 웹 자료를 수집하는 양적 성장을 보였다면, 이용자가 납득하고 동의할 만한 질적 성장도 아울러 보여주어야 마땅할 것임에도 이 연구에서 살펴본 공개된 콘텐츠의 품질은 의심스러운 수준이다. 그렇다면 공개되지 않은 99%의 웹 자료들의 수준은 과연 어떨지, 철저한 검증이 필요하다. 우선은 이 연구에서 지적된 수집 콘텐츠의 주제 분포의 편중, 웹사이트 저작자/발행자 권위 및 학술적 가치의 근거 미약, 아카이빙 대상의 정의 부적합, 수집 및 이용 통계 처리의 모호성 등 같은 문제점만이라도 개선되기를 기대할 뿐이다. 국립중앙도서관 자체가 이러한 문제점을 인식하고 포괄적이며 체계적인 개선방안을 시급히 마련해야 할 것이다.

참고문헌

국립중앙도서관. 2009. 『2008 국립중앙도서관 연보』. 서울: 국립중앙도서관.

국립중앙도서관. 2011. 『2010 국립중앙도서관 연보』. 서울: 국립중앙도서관.

국립중앙도서관, OASIS. 2009a. OASIS 소개. 개요. [인용 2009. 5. 22]. 〈http://www.oasis.go.kr/intro/intro_overview.jsp〉.

국립중앙도서관, OASIS. 2009b. OASIS 소개. 수집보존프로세스. [인용 2009. 5. 30].

〈http://www.oasis.go.kr/intro/intro_gatherprss.jsp〉.

국립중앙도서관, OASIS. 2009c. OASIS 소개. 추진경과. [인용 2009. 5. 22]. 〈http://www.oasis.go.kr/intro/intro_history.jsp〉.

국립중앙도서관, OASIS. 2009d. OASIS 소개. 자원 수집 지침.

국립중앙도서관, OASIS 홈페이지. [인용 2011. 5. 25]. 〈http://www.oasis.go.kr〉.

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

김유승. 2008. 복합적 웹 아카이빙 정책에 관한 고찰: 프랑스국립도서관의 사례를 중심으로. 『한국문헌정보학회지』, 42(4): 159-179.

서혜란. 2004. 『디지털 납본제도 방안』. 2004년 디지털 유산보존에 관한 기초연구 용역논문. 서울: 국립중앙도서관.

윤정옥. 2010. 웹 아카이브 OASIS에 관한 고찰. 『한국문헌정보학회지』, 44(2): 5-27.

이소연. 2008. 믿을 수 있는 디지털 아카이브 인증기준: OASIS 적용사례. 『情報管理學會誌』, 25(3): 5-25.

이혜원. 2004. 『온라인 디지털 자원 구축 사례: 국립중앙도서관을 중심으로』. 2004년 디지털 유산보존에 관한 기초연구 용역논문. 서울: 국립중앙도서관.

최원태. 2005. 『디지털유산 보존과 국가대표도서관의 정책방안』. 서울: 국립중앙도서관.

Library of Congress Web Archives. Minerva. 2009. "Minerva." [cited 2009. 5. 30]. 〈http://lcweb2.loc.gov/diglib/lcwa/html〉/lcwa-home.html〉.

Library of Congress Web Archives. Minerva. 2011a. "United States Election 2008 Web Archive." [cited 2011. 6. 3]. 〈http://lcweb2.loc.gov/diglib/lcwa/html/elec2008/elec2008-overview.html〉.

Library of Congress Web Archives. Minerva. 2011b. [cited 2011. 6. 3]. 〈http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html〉.

National Library of Australia and Partners. 2009a. "PANDORA. Australia's Web Archive. PANDORA Overview." [cited 2009. 5. 30]. 〈http://pandora.nla.gov.au/overview.html〉.

National Library of Australia and Partners. 2009b. "PANDORA. Australia's Web Archive. PANDORA Archive Size and Monthly Growth." [cited 2009. 5. 30]. 〈http://pandora.nla.gov.au/statistics.html〉.

National Library of Australia and Partners. 2009c. "PANDORA. Australia's Web Archive. Selection Guidelines." [cited

2009. 5. 30].
〈http://pandora.nla.gov.au/guidelines.html〉.

National Library of Australia. PANDORA, Australia's Web Archive. 2011a. "PANDORA Archive Size and Monthly Statistics." [cited 2011. 6. 6]. 〈http://pandora.nla.gov.au/statistics.html〉.

National Library of Australia. PANDORA, Australia's Web Archive. 2011b. "Statistics for PANDORA." [cited 2011. 6. 6]. 〈http://stats.nla.gov.au/_reports/pandora/monthly/5-2011/awstats.pandora.html〉.