웹 아카이브 OASIS 수집 콘텐츠의 분석

An Analysis of the Contents of OASIS, the National Web Archive in Korea

윤 정 옥(Cheong-Ok Yoon)*

<목 차>

Ⅰ. 머리말
Ⅱ. 선행 연구
Ⅲ. OASIS 수집 웹 자원의 현황 분석
1. OASIS 웹 자원의 수집 실적
2. 수집 웹사이트의 수량적 성장
3. 수집 웹사이트의 주제별 분포
4. 수집 웹사이트의 지속성
5. 수집 웹사이트의 최신성
6. 수집 웹사이트 통계의 정확성
Ⅳ. 맺음말

초 록

이 연구에서는 국립중앙도서관이 구축 및 운영하는 웹 아카이브 OASIS 콘텐츠의 특성과 현황을 살펴보았다. 2013년 12월-2014년 11월 OASIS에 공개된 웹사이트 55,581건의 수량적 성장과 주제 분포, '최신 수집자료'의 아카이빙 현황 등을 검토하였다. 급격한 수량적 성장에도 불구하고, '사회과학'(63.6%)에 집중한 주제 편향성, '정치학'(34.7%. 2003년 전체의 21.4%)의 과도한 편중, '최신 수집자료'의 저작자 권위 및 학술적 가치의 근거 미약, 웹사이트와 인스턴스의 혼용에 따른 통계의 중복 및 부정확성 등 문제점이 다시 확인되었다. 양적 성장에 동반하지 않는 질적 수준 문제가 지속되며, 시급한 수집정책 개선과 품질제어가 필요한 것으로 나타났다.

키워드: 오아시스, 웹 아카이브, 국립중앙도서관, 디지털 자원, 웹 자원 보존

ABSTRACT

The purpose of this research is to examine the characteristics and current status of OASIS, a web archive, developed and operated by the National Library of Korea. From December 2013 to November 2014, an analysis of a numerical growth and subject distribution of 55,581 websites archived at OASIS shows many problems in quality, including an overwhelming proportion of 'Social Science' and its subclass 'Politics', consequential lack of balance in subject distribution, lack of authority or scholarly value of some contents, unclear application of selection criteria for personal creators/publishers, and inaccurate and overlapping statistics. Despite an impressive growth in quantity, immediate improvement of selection policies and quality control is needed.

Keywords: OASIS, Web archive, National library of Korea, Digital resources, Preservation of web resources

* 청주대학교 문헌정보학과 교수(jade@cju.ac.kr)
•논문접수: 2014년 11월 25일 •최초심사: 2014년 11월 25일 •게재확정: 2014년 12월 22일
•한국도서관·정보학회지 45(4), 45-65, 2014. [http://dx.doi.org/10.16981/kliss.45.201412.45]

Ⅰ. 머리말

지난 1990년대 중반부터 웹상에서 생산되고 유통되는 정보자원의 규모는 기하급수적으로 커졌다. 그동안 여러 나라에서 급격히 증대하는 웹 정보자원 가운데 가치 있는 자국 관련 인터넷 자료를 다양한 기준과 관점에 의거하여 수집, 보존하고 이용자에게 제공하려는 목적으로 국가 웹 아카이브를 구축하기 시작하였고, 지난 십여 년 사이 그러한 노력의 성과 또한 웹상에서 공개되고 있다. 호주는 일찍이 1996년부터 National Library of Australia(2014)가 "호주와 호주인에 관련된 역사적 온라인 간행물의 컬렉션"인 PANDORA를 구축하고 있으며, 미국의회도서관도 일찍부터 "미래 세대가 이용할 수 있도록 현재의 디지털 표현물(digital expressions)을 보존"하는 것의 중요성을 강조하면서 국가 디지털 기반구조 및 보존 프로그램(National Digital Infrastructure and Preservation Program)의 계획을 선도하였다(Beagrie 2003).영국은 2004년부터 British Library가 UK Web Archive(2014)를 구축하여 "연구를 공개하고, 영국 전역의 생활, 관심 및 활동의 다양성을 반영하며 웹 혁신을 증명하는 웹사이트" 및 "브리핑, 보고서, 정책선언, 그 밖의 단명하지만 중요한 형태의 정보를 수록하는 회색문헌"을 수집하고 있다.

국가 웹 아카이브 구축의 목적은 UK Web Archive(2014)가 "잠재적인 '디지털 블랙 홀'이라는 도전에 대응"하고, 가능한 한 많은 웹사이트들을 보존하고 "미래 세대를 위해 핵심적인 UK 웹사이트에 대한 영구적인 온라인 접근을 제공"한다고 한 데서 잘 요약되고 있다. 국가 웹 아카이브를 위해 단명하고, 급격히 소멸하는 웹상의 자료들 가운데 보존할 만한 가치를 가진 자원을 가려내는 것(Day 2003)은 막대한 시간과 인적 자원을 요구하는 작업이다. 따라서 주요한 국가 웹 아카이브들은 다양한 정책 방향을 채택하여 미국의 MINERVA처럼 선별적 주제 컬렉션을 구축하거나, 호주나 영국처럼 단일 기관의 포괄적, 망라적 수집보다는 국가대표도서관과 복수의 유관 기관들의 분담 및 협력적 수집을 수행하기도 한다. 지난 십여 년 사이 웹 자원의 수집과 보존 기술과 방법은 매우 발전해 왔다. 하지만, 다양한 방법과 매체를 통해 생산되고 배포되는 웹 자원의 분산, 다양한 기술, 하드웨어와 소프트웨어의 급격한 등장과 소멸 등은 여전한 문제로서 최대한 조기 수집과 장기적 보존을 위한 변환 등이 지속적 과제로 여겨진다. 특히 특정 주제와 관련하여 새롭게 등장하고 빠르게 변화하는 콘텐츠의 발견과 아카이빙, 또한 아카이빙을 자극하는 동향을 추적하는 알고리즘의 발전 같은 개별적 과제들도 중요시되고 있다(Meyer 2011).

우리나라는 2004년 국립중앙도서관이 OASIS(Online Archiving & Searching Internet Sources)라는 명칭의 국가 웹 아카이브를 구축하고 "가치 있는 인터넷 자료를 국가적인 차

원에서 수집 · 축적하여 미래 세대에 연구 자료로 제공"하겠다는 목표를 선언하였다(국립중앙도서관. OASIS 2009). OASIS가 출범한 지 벌써 십 년에 이른 만큼 수집 웹 자원의 수량적 성장과 더불어 질적인 성장 현황에 대한 체계적인 점검이 필요한 시점이다. 국가 웹 아카이브의 선두주자라 할 만한 호주의 PANDORA도 운영 과정에서 계속적 점검과 보완을 거치며 발전하였고(Crook 2008), 영국의 UK Web Archive는 선별적 수집으로 시작되었지만, 새로운 법령의 시행에 따라 망라적 수집으로 확대되는 변화를 겪은 것(British Library 2013)처럼 어떤 기관이든 현재 상황의 분석과 이해에 기반하여 미래의 지향점을 확인 혹은 재설정할 필요가 있다. OASIS도 예외는 아니다.

그동안 OASIS의 현황과 발전에 주목해온 일련의 연구들(김유승 2007, 2008; 윤정옥 2010, 2011, 노영희, 고영선 2012)에 따르면 OASIS는 공개된 웹 자료의 선정기준 미흡, 수집 통계의 부정확성, 수집된 콘텐츠 가치의 적절성 부족 등 다수의 문제점을 노출하고 있으며, 수집 기준과 정책의 개선이 시급한 것으로 나타났다. 이 연구는 이전 연구들의 연속선상에서 2014년 11월 시점의 OASIS 콘텐츠 현황을 검토하여, 문제점을 도출하고 가능한 개선의 방향을 제안하는 것을 그 목적으로 하였다.

이 연구에서는 국가 웹 아카이브 관련 국내외 문헌 및 웹사이트 등을 검토하고, 2013년 12월 26일부터 2014년 11월 10일 사이 수시로 OASIS 홈페이지에 공개된 웹 자원의 현황을 분석하였다. 우선 2009년, 2013년 및 2014년 OASIS 수집 웹 자원 통계를 중심으로 주제별 수량적 성장의 추이를 분석하고, 2014년 11월 시점의 '최신 수집자료' 웹사이트와 '주제 브라우징' 및 '테마 브라우징' 리스트에서 개별 웹사이트와 아카이빙 버전을 검토하였다.

Ⅱ. 선행 연구

웹 아카이빙 자체에 대한 연구는 일찍부터 많이 이루어졌으나, 실제 국가 웹 아카이브를 구축하고 운영한 경험으로서 호주 National Library of Australia의 Web Archiving & Digital Preservation Branch의 Crook의 연구(2008)가 특히 주목할 만하다. Crook은 출범 10여 년이 지난 PANDORA의 온라인 자료의 수집 범위와 방법의 변화, 새로운 아카이빙 기술의 발전 및 도서관의 적응성 등을 분석하였다. 초기에 아카이빙은 인터넷 상 자료들 중 선별적이며 아주 작은 부분만 수집할 수 있었으나, 오늘날 급격한 기술 발전과 정책 변화로 웹 상에서 이용할 수 있는 것과 아카이빙 할 수 있는 것 사이 격차가 줄어들고 있음을 강조하며, PANDORA 운영 기간 동안 자체의 디지털 아카이빙 시스템인 PANDAS가 3차례나 전면적 변화를 겪었음을 보고하였다. 이 연구는 초기에 꼼꼼한 정책적 및 기술적 검토와 준비로 시작

된 국가 웹 아카이브의 모델이라고도 할 수 있는 PANDORA도 실제 운영과 실행의 과정에서 시행착오와 궤도 수정을 경험해야 한다는 것을 증명하였다. 한편 앞에서 언급했던 호주의 PADI 이니셔티브는 주요한 웹 아카이브 관련 연구 및 실행 사례들을 분석하여 자료를 제공하고 있다. 우리나라 OASIS 같이 비교적 후발 주자인 국가 웹 아카이브들은 이러한 선행 모범사례 및 실패담을 동시에 참조하여 시작할 수 있다는 점에서 오히려 유리한 입장에 선다고도 할 수 있다. 한편 Toyoda와 Kitsuregawa(2012)는 주로 주로 기술적 발전에 주목하며, International Internet Preserving Consortium (IIPC)이 2008년 Internet Archive가 이전에 사용하였던 ARC 파일 포맷에 기반한 Web ARChive(WARC) 파일 포맷을 확정하였고, 그밖에도 데이터 수집, 컬렉션 저장소 및 관리를 위한 다양한 툴 키트를 제공하고 있다는 점 등을 강조하고 있다(Toyoda and Kitsuregawa 2012).

포르투갈의 Foundation for National Scientific Computing(FCCN)의 Gomes, Miranda와 Costa의 연구(2011)에 따르면 2011년 현재 세계 26개국에 42개 웹 아카이빙 이니셔티브가 있다. 연구자들은 OECD 34개국 중 21개국(62%)이 적어도 1개의 웹 아카이빙 이니셔티브를 주관하는 것은 선진국들이 웹 아카이빙을 중요시한다는 증거라고 강조하였다. 아울러 호주 National Library of Australia의 Preserving Access to Digital Information(PADI) 이니셔티브(National Library of Australia 2014)는 세계 여러 나라의 주요한 아카이빙 이니셔티브 정보를 제공하며 잘 확립된 국가 수준의 웹 아카이빙 프로그램을 갖고 있거나 관련 이슈를 적극적으로 추구하는 나라들로 호주, 오스트리아 등 17개국을 들고 있는데, 이들 중 리투아니아를 제외한 나머지 나라들이 모두 앞서 Gomes, Miranda와 Costa(2011)가 지적한 것처럼 OECD 회원국가라는 점은 주목할 만하다. 한국의 OASIS는 여기 포함되지 못하였다.

국내에서는 2004년 OASIS 출범을 전후하여 국가적 웹 아카이브의 필요성, 정책적 목표, 전략 개발 등 측면을 분석하고 전망한 연구들이 등장하였다(서혜란 2004; 이혜원 2004). 이후 OASIS가 디지털 아카이브로서 웹 자원의 수집, 보존, 관리 등 전 단계에서 갖출 조건을 다룬 이소연의 연구(2008), OASIS 수집 방법의 문제점과 서비스 활용 방안을 논한 김유승의 연구(2007), OASIS의 광범하고 심층적인 수집 범위 설정, 다양한 아카이빙 정책 적용 및 타 기관과의 협력적 웹 아카이빙 전략 수립 등 필요성을 논한 김유승의 후속 연구(2008) 등이 나왔다.

윤정옥(2010)은 OASIS 서비스를 통해 제공되는 웹 자원의 콘텐츠와 검색 관련 문제점에 주목하여, 2009년 5월부터 7월 사이 '문학'과 '사회과학' 분야 콘텐츠 및 이용 가능한 서비스를 검토하였다. 수집 콘텐츠의 주제 분포의 편중, 저작자/발행자 편향성과 권위의 근거 미약, 정보의 유일성과 최신성 결여, 웹 문서와 웹사이트의 중복 수집, 학술적 가치의 근거 결여 등 문제점을 발견하여, 주제전문가의 활용과 실명제, 메타데이터 요소 추가 및 기본적 목표의식의 지속적 확인 등이 필요함을 지적하였다. 윤정옥(2011)은 다시 2011년 5월 OASIS 주제

별 디렉토리에서 제공되는 '철학', '종교' 등 5개 주제 및 '사회과학'과 '기술과학' 소주제의 수집 웹사이트 55건을 검토하고, '최신 자료'와 '많이 본 자료' 리스트를 분석하여, 콘텐츠와 서비스 상 동일한 문제가 여전히 존재하며 개선이 시급함을 강조하였다. 노영희와 고영선(2012)은 영국의 National Archives, 미국의 MINERVA, 호주의 PANDORA 등 국내외 주요한 웹 아카이빙 선정지침을 분석하고, OASIS의 웹 자료 선정지침 개선이 필요함을 지적하였다. 이들은 웹 자료의 범위 및 용어 정의, 수집의 기본원칙, 방법, 주기 등의 개선, 선정 제외 자료에 대한 구체적 지침을 제안하였다.

Ⅲ. OASIS 수집 웹 자원의 현황 분석

여기에서는 지난 2013년 12월부터 2014년 11월 사이 수시로 OASIS 홈페이지에 공개된 웹 자원 현황을 검토한 결과를 토대로 하여, OASIS 수집 자료의 수량적 성장, 주제별 분포 추이, 콘텐츠의 지속성과 최신성 등을 분석하였다.

1. OASIS 웹 자원의 수집 실적

<표 1> 연도별 웹 자원 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
구분	-2010	2011	2012	-	-	-	-	합계
웹 자원	267,682	183,487	80,504	-	-	-	-	531,673

국립중앙도서관은 2004년 처음 OASIS가 40,096건의 웹 자료(웹사이트와 웹문서)를 수집한 이래로 매년 수집 실적을 공개하였다. <표 1> 상단 두 줄에서 보는 2004년부터 2010년까지 공개된 수집 웹 자료의 규모는 『2010 국립중앙도서관연보』에 공개된 것으로 꾸준한 수량적 성장세를 보여주었다. 이 연보에 따르면 2010년 말까지 누적된 수집 웹 자료의 총수는 모두 584,565건이었다(국립중앙도서관 2011, 68). 한편 <표 1> 하단 두 줄의 연도별 수치는 최근 『2012 국립중앙도서관연보』 (국립중앙도서관 2013, 76)에 공개된 것이다. 이 연보에서는 웹 자료가 아니라 웹 자원이라는 용어를 사용했고, 2012년 웹 문서 65,284건, 웹사이트 15,220건, 총 80,504건을 수집한 것으로 나타났다. 이 연보에 수록된 <표 4-1> '연도별 웹 자원 수집실적'에는 웹 자원(웹사이트, 웹 문서)의 건수는 '-2010년' 267,682건, 2011년 183,487건, 2012년 80,504건으로 합계는 531,673건이라고 하였다.

<표 1> 상단과 하단의 누적 수치는 큰 차이가 있다. 『2010 국립중앙도서관연보』에서 2010년까지 웹 자료 누적 건수가 584,565건, 『2012 국립중앙도서관연보』 에서는 2010년 당시 웹 자원 누적 건수가 267,682건이라고 하여, 32만여 건이나 차이가 난다. 이러한 통계 차이는 웹 자원과 웹 자료라는 용어를 혼용한 데서 비롯된 것이다. OASIS에서 이 용어들을 구체적으로 설명하지 않았으나 현재 OASIS 검색 결과 화면에서 '웹사이트 검색 결과', '웹페이지 검색 결과' 및 '웹자료 검색 결과'로 제시하는 데서 그 정의를 볼 수 있다. 이 화면의 설명에 따르면, '웹사이트 검색 결과'는 '국립중앙도서관에서 수집한 웹사이트 보존파일'을 보여주고, '웹페이지 검색 결과'는 '웹사이트 내 웹페이지의 내용'을 보여준다. 또한 '웹자료 검색 결과'는 '웹페이지 내에서 제공하는 문서(pdf, hwp 등)'를 보여준다. 즉 웹 자원의 수준을 웹사이트, 웹페이지, 웹 자료의 삼단계로 구분하였고, 웹 자료는 최하위의 웹 문서를 나타내는 것이다. 따라서 『2010 국립중앙도서관 연보』 는 웹 자료라는 하위 수준 단위로, 『2012 국립중앙도서관 연보』 는 웹 자원이라는 상위 수준 단위로 통계를 제시함으로써 그러한 수치 차이가 나타난 것이라 할 수 있다.

2. 수집 웹사이트의 수량적 성장

<표 2>는 2009년 10월부터 2014년 11월 현재까지 OASIS에 수집된 웹사이트의 수량적 성장 현황을 보여주고 있다. OASIS는 '주제별 브라우징' 리스트에 KDC에 따라 웹사이트를 10개 주제로 구분하여 통계 수치와 웹사이트 리스트를 제공한다.

<표 2> OASIS 수집 웹사이트의 수량적 변화(2009-2014)

주제	2009.10.9		2013.12.26		2009-2013	2014.11.9		2013-14
	건수	비율	건수	비율	증가(배)	건수	비율	증감(건)
총류	31	12.0%	1,919	4.2%	61.9	1,817	3.3%	-102
철학	3	1.2%	239	0.5%	79.7	349	0.6%	110
종교	20	7.7%	1,834	4.0%	91.7	1,898	3.4%	64
사회과학	78	30.1%	28,373	61.7%	363.8	35,336	63.6%	6,963
순수과학	6	2.3%	956	2.1%	159.3	1,054	1.9%	98
기술과학	80	30.9%	6,082	13.2%	76.0	6,696	12.0%	614
예술	15	5.8%	4,773	10.4%	318.2	5,635	10.1%	862
언어	5	1.9%	264	0.6%	52.8	509	0.9%	245
문학	11	4.2%	442	1.0%	40.2	456	0.8%	14
역사	10	3.9%	1,104	2.4%	110.4	1,831	3.3%	727
합계	259	100%	45,986	100%	135.4	55,581	100%	9,595

<표 2>는 2013년 12월 26일과 2014년 11월 9일 두 차례 OASIS의 '주제별 브라우징' 리스트에 공개된 웹사이트 통계를 지난 2009년 10월 9일 당시 공개되었던 웹사이트 통계와 비교하여 구성한 것이다. 2009년 데이터는 이전 연구(윤정옥 2011)를 근거로 구성한 것으로 당시 OASIS는 49,441건의 웹 문서와 259종의 웹사이트를 구별하여 공개하였으나, 2014년 현재에는 웹사이트 통계만 볼 수 있다. OASIS는 현재 '건'을 계량단위로 쓰고 있으므로, 이 연구에서도 이를 사용하였다. 2014년 11월 시점에 OASIS에 공개된 웹사이트는 모두 55,581건으로 2013년 12월 당시 45,986건보다 9,595건 늘어난 규모로 월 평균 90여 건 증가한 수준이다. 2009년 259건, 2011년 350건을 수집했던 것에 비하면 엄청난 증가율을 보이고 있다.

<표 2>에서는 2009년, 2013년 및 2014년 현재 수집 웹사이트의 주제 분야별 규모 및 분포의 변화도 알 수 있다. 2009년 당시 수집 및 공개된 웹사이트 수가 워낙 적었으므로 2013년까지 불과 4년 사이에 평균 135.4배 증가하였다. 모든 주제에서 수집 규모가 급격히 증가하였고, 특히 '사회과학'(363.8배)과 '예술'(318.2배) 분야의 급증은 크게 주목할 만하다.

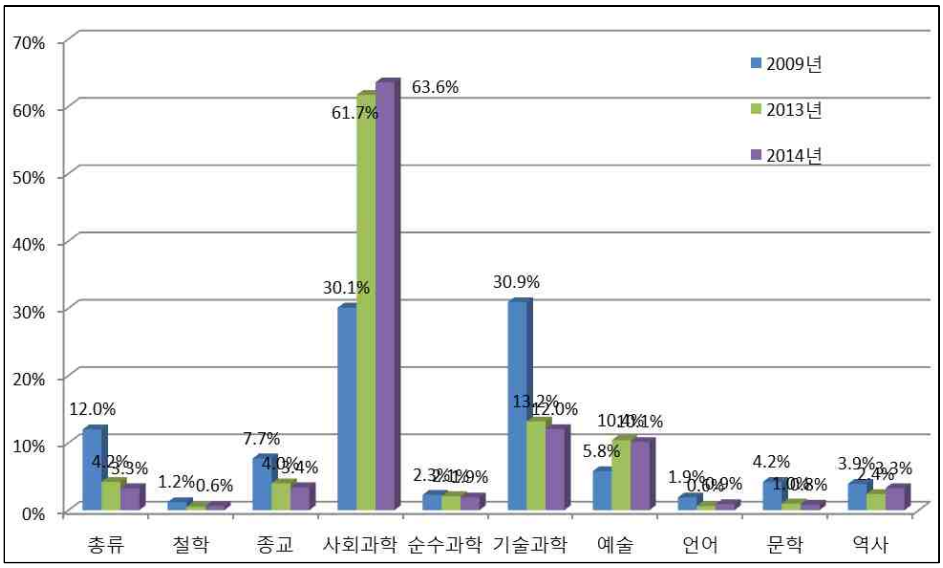
여기에서 또한 눈여겨 볼 것은 2013년부터 2014년 사이 수집 건수가 급격히 증감한 총류, 사회과학, 기술과학, 예술 및 역사 분야이다. 이 기간 동안 사회과학은 6,963건 증가하였고, 기술과학은 614건, 예술은 862건, 역사는 727건 각각 증가하였다. 특히 사회과학(63.6%), 기술과학(12.0%) 및 예술(10.1%) 분야는 2014년 수집 웹사이트 전체에서 가장 주제 분포도가 높은 상위 3개 분야로, 합쳐서 85.7%에 이른다. 2013년 당시 이 세 주제가 전체 85.3%를 구성했던 것보다 약간 더 비중이 늘어났다.

총류는 2013년 1,919건에서 2014년 1,817건으로 102건 감소하여 유일하게 감소한 분야이다. 하지만 단순한 수치 증감만으로 감소 이유와 실제 현황은 알 수 없다. 이 분야에서는 새로운 웹사이트가 전혀 수집되지 않은 것인지, 아니면 새로운 것을 수집하였어도 삭제한 웹사이트 수가 더 많았던 것인지 알기 어렵다. 만약 OASIS에서 상시 혹은 수시로 이전 수집 자료를 모니터 하고 적절하지 않은 웹사이트가 있을 경우 삭제하는 작업을 하고 있다면 신규 수집 건수와 더불어 폐기 혹은 삭제 통계를 제시해야 수치의 증감이 설명될 수 있다.

3. 수집 웹사이트의 주제별 분포

가. 전체 웹사이트의 주제별 분포

<그림 1>은 2009년부터 2014년 사이 OASIS 수집 웹사이트의 주제별 분포 변동을 보여준다. 앞 절의 <표 2>와 <그림 1>을 함께 보면 수집 웹사이트들의 주제별 분포가 매우 불균형하며, 지속적 불균형 상태가 개선되지 않고 있음을 알 수 있다.



<그림 1> OASIS 수집 웹사이트의 주제별 분포 변동: 2009년-2014년

가장 눈에 띄는 것은 앞 절에서 언급한 대로 2014년 현재 '사회과학' 웹사이트가 전체의 63.6%에 달하며 2013년의 61.7%보다도 증가하였다는 점이다. '사회과학'은 지난 2009년 공개된 웹사이트 259건 가운데 78건(30.1%)으로, 30.9%(80건)이었던 '기술과학' 웹사이트와 비슷하게 큰 비중을 차지하였다. 전체 공개 건수가 많지 않았던 2009년에도 비중이 컸지만 2013년과 2014년 시점에서는 '사회과학' 한 분야만이 압도적일 정도로 주제 편중이 더욱 심화되었다. 또한 2009년 5.8%(15건)에서 2014년 10.1%(5,635건)로 증가한 '예술' 분야를 제외하고는, 이 기간 동안 다른 8개 주제 웹사이트 비율은 모두 감소하였다. 특히 '기술과학'은 전체의 12.0%(6,696건)로서, 2009년에 비하여 수집 건수는 크게 증가하였으나 주제 비중은 현저히 줄어들었다.

2014년 '총류'(1,817건, 3.3%), '종교'(1,898건, 3.4%), '역사'(1,831건, 3.3%)의 3개 주제는 가까스로 3%를 넘어섰으나, '철학'(349건, 0.6%), '언어'(509건, 0.9%), '문학'(456건, 0.8%)의 3개 주제는 1%도 넘지 못하였다. 이처럼 2009년에서 2014년 사이 수집 웹사이트 규모는 엄청나게 증대하였으나, 주제의 불균형은 매우 심화되었다.

물론 이러한 주제의 불균형은 OASIS 수집 정책의 문제가 아니라 국내에서 생산되고 유포되는 웹사이트들의 주제 분포의 불균형에 원인이 있을 수도 있으나 확인하기 어렵다. 통계청 통계포털에 따르면 2012년 kr 도메인을 가진 국내 웹사이트는 1,097,557개(통계청 사용 단위)에 이른다(통계청 2012). 이 통계는 .com, .org, .net 등 도메인을 가진 국내 웹사이트들은 포함하지 않기 때문에 실제 전체 현황을 나타내지 못한다. 더욱이 웹사이트의 주제별 분포

도는 알 수 없다. 따라서 OASIS 수집 웹사이트의 주제 불균형의 직접적 원인을 파악하기 어렵지만, 수집 기준이나 절차 자체에서 불균형이 기인한 것은 아닌지, 얼마나 그런 추세가 지속될 지 계속 관찰하고 원인을 파악하고자 노력할 필요가 있다.

나. '사회과학' 웹사이트의 분포

<표 3> '사회과학' 웹사이트의 소주제 분포 변동: 2011년-2013년

	2011.6.7		2013.12.27		2011-13
소주제	건수	비율	건수	비율	건수 증가
사회과학	8	7.2%	426	1.5%	53.3
경제학	25	22.5%	4,759	16.8%	190.4
정치학	8	7.2%	9,835	34.7%	1229.4
법학	8	7.2%	359	1.3%	44.9
풍속, 예절, 민속학	21	18.9%	415	1.5%	19.8
통계학	0	0.0%	14	0.0%	0.0
사회학, 사회문제	20	18.0%	4,710	16.6%	235.5
행정학	7	6.3%	2,798	9.9%	399.7
교육학	10	9.0%	4,899	17.3%	489.9
국방, 군사학	4	3.6%	158	0.6%	39.5
총수	111건	100%	28,373건	100%	255.6배

<표 3>은 전체 수집 웹사이트 중 가장 큰 비중을 차지하는 '사회과학' 주제 분야의2011년에서 2013년 사이 소주제 분포 변동을 보여준다. 2013년에는 OASIS 웹사이트 '주제별 브라우징' 리스트에서 KDC 10개 주류의 소주제까지 통계가 공개되었기 때문에 <표 3>의 정리가 가능하였으나, 2014년 11월 현재 더 이상 소주제 통계는 공개되지 않는다. '사회과학' 웹사이트는 2009년 78건에서 2011년 111건으로 소량 증가하였으나, 2013년 12월 말 당시 28,373건으로 급격히 증가하였으며, 전체 웹사이트의 61.7%를 구성하였다. 2011년 6월에서 2013년 12월 사이 수집 총 건수는 무려 255.6배 증가한 것이다.

이 기간 동안 '사회과학' 10개 소주제의 분야별 웹사이트 건수와 분포 변화를 살펴보면, '사회과학'은 8건(7.2%)에서 426건(1.5%)으로, '경제학'은 25건(22.5%)에서 4,759건(16.8%)으로, '정치학'은 8건(7.2%)에서 9,835건(34.7%)으로, '법학'은 8건(7.2%)에서 359건(1.3%)으로, '풍속, 예절, 민속학'은 21건(18.9%)에서 415건(1.5%)으로 각각 증가하였다. '통계학'은 2011년 당시 1건도 수집되지 않았으나 14건(0%)이 수집되었고, '사회학, 사회문제'는 20건(18.0%)에서 4,710건(16.6%)으로, '행정학'은 7건(6.3%)에서 2,798건(9.9%)으로, 교육학은 10건(9.0%)에서 4,899건(17.3%)으로, '국방, 군사학'은 4건

(3.6%)에서 158건(0.6%)으로 모두 증가하였다. '통계학'은 수집 건수가 워낙 미미하여 증가 비율조차 0%로 머물고 있으나, 다른 분야들은 수십 배에서 수백 배 증가하였고, 특히 정치학 분야는 엄청난 규모로 증가하였다.

'정치학' 분야 웹사이트는 2011년부터 2013년까지 무려 1229.4배 증가하였고, 전체 '사회과학' 내에서 분포 비율도 7.2%에서 34.7%로 급증하였다. 당시 OASIS 수집 웹사이트 45,986건 중에서 '사회과학-정치학' 주제 웹사이트는 9,835건으로 전체의 21.4%를 차지하는 규모이다. 그밖에 '경제학', '사회학, 사회문제', '행정학', 및 '교육학'의 4개 소주제 분야는 각각 10% 내외의 분포 비율을 보였지만, '법학' 등 나머지 5개 소주제는 다 합쳐도 4.9% 밖에 되지 않았다. 전체 웹사이트 가운데 '사회과학'의 비중이 압도적인 상황에서, 그 안의 소주제들 또한 심한 분포 불균형을 보이고 있다.

다. 전체 웹사이트의 테마별 분포

OASIS는 '주제별 브라우징' 리스트와 더불어 '테마별 브라우징' 리스트도 공개하고 있다. 2014년 11월 11일 현재 '테마별 브라우징'은 다음과 같은 8개 테마의 통계를 제시한다(괄호 안은 건수): 국가회의(0), 기관(0), 선거(883), 정부부처(482), 학회 및 연구기관(2607), 기타(21841), 행사(430), 기업(6).

여기 공개된 웹사이트 건수는 모두 26,249건으로 앞 절에서 살펴본 '주제별 브라우징' 리스트의 공개 웹사이트 55,581건의 47.2% 정도이다. 이들 8개 테마 중 '국가회의'와 '기관' 테마로 분류된 웹사이트는 0건이며, '기타'가 21,841건으로 83.2%에 이르러, 사실상 테마 분류는 무의미하게 보인다. 그럼에도 불구하고 테마 분류의 내용을 검토하고자 가장 규모가 작은 테마 '기업'(6건)을 브라우징한 결과는 다음과 같았다:

1. 반여중학교 (URL: http://banyo.ms.kr) 사회과학 > 교육학 > 중등교육
2. 퍼스트 스텝스 (URL: http://www2.firststepscanada.org/) 사회과학 > 사회학, 사회문제 > 사회단체
3. 울산현대산악마라톤대회 (URL: http://www.hdsanak.com/) 예술 > 오락, 스포츠 > 육상경기
4. 유비쿼터스응용연구실 (URL: http://caeagle.yonsei.ac.kr) 총류 > 컴퓨터과학
5. 제주문화포럼 (URL: http://munhwaforum.or.kr/) 역사 > 아시아(아세아) > 한국
6. 신아시아연구소 (URL: http://nari.re.kr/) 역사 > 아시아(아세아)

상기한 6개의 웹사이트가 테마 '기업'에 속한다고 하기는 어렵다. 만약 레코드 1. '반여중학

교'가 '기업'이라면, 수집 웹사이트 '키워드 검색'에서 '중학교'를 입력하면 나오는 476개 중학교도 같은 테마로 분류되어야 할 것이다. 초록 등 상세정보를 검토해도 이 특정한 중학교가 '기업'으로 분류된 이유는 찾을 수 없다.

다른 테마로 분류된 웹사이트들도 마찬가지다. '행사' 테마로 분류된 430건의 브라우징 결과는 국제울진대게축제 등 축제, 엑스포, 페스티벌, 대회, 포럼을 망라하는 비교적 다양한 웹사이트들을 보여주었다. 동시에 한국문화예술위원회 온라인기부, 한국노화예방연구원, 흑자경영연구소 등 다소 이해하기 어려운 웹사이트들, 다시 말하면 '행사' 테마에 적합하지 않은 것들이 다수 포함되어 있었다. '정부부처' 테마의 482건 또한 각급 정부 부처와 관련된 도서관, 교육청, 주민센터, 총영사관 등을 포함하지만, 무대예술전문인 자격검정위원회, 대한의사협회 의학용어위원회, 대한민국 베트남참전 유공전우회 등 모호한 단체들도 포함하고 있다.

이처럼 특정 테마의 웹사이트 분류 기준이 불확실하거나 부정확한 한편 망라성 또한 문제이다. 예를 들어, '수집 웹사이트' 55,581건의 리스트에서 주민센터는 512건이 검색되는데, 이 '정부부처' 테마 리스트에서는 용두동 동주민센터 등 175건만 나온다. 도서관 또한 '수집 웹사이트' 리스트에서는 1,044건이 검색되는데 여기에서는 양주시립도서관 등 43건만 검색된다. 이러한 '테마별 브라우징' 리스트의 현황은 기왕에 수집한 자료를 어떻게 조직하여 이용자로 하여금 접근할 수 있게 하는가, 즉 수집 이후 부가가치적 처리에 관련되는데, 수집 웹사이트의 부분적 공개, 부적합한 테마 분류 등 여러 가지 문제점을 내보이고 있다.

4. 수집 웹사이트의 지속성

OASIS가 가치 있는 웹 자원을 미래 세대를 위하여 수집하고 축적한다면, 그동안 수집한 자원 또한 지속적으로 잘 보존 및 관리해야 한다는 기대가 있다. OASIS는 2014년 11월 현재 홈페이지 '주제별 브라우징' 디렉토리에 웹사이트 55,581건을 공개하고 있어, 이용자가 접근 가능한 웹사이트 규모는 이전보다 상당히 늘어났다.

여기에서는 이전에 수집된 웹사이트의 현재 보존 여부 및 접근 가능성을 보기 위하여 이전 연구(2011)에서 살펴보았던 55건을 검토하였다. 사이트명 키워드 검색 결과, 2014년 3월 시점에 이들 중 30건(54.5%)이 아카이빙 되어 있었고, 다른 25건(45.5%)은 검색이 되지 않았다. 이들을 최초 아카이빙 연도를 기준으로 보면 웹사이트의 장기적 보존 기준 혹은 지속성에 의문을 갖게 된다. 당시 보존되었던 55건 가운데 최초 아카이빙 연도를 알 수 없었던 대한철학회 등 4건은 현재도 모두 아카이빙 되어있다. 그러나 최초 아카이빙 시점이 2004년인 '팬코리아 영어교육학회' 등 7건 중 5건(71.4%), 2005년 아카이빙 된 '경북대 김문기교수와 함께 하는 한국고전의 세계' 등 34건 중 13건(38.2%), 2006년 아카이빙 된 한국생물

정보학회 홈페이지 등 6건 중 3건(50.0%)이 현재 남아있다. 그밖에 2007년과 2010년 처음 아카이빙 된 웹사이트도 각각 2건씩 남아있다.

최초 아카이빙 시점이 2004년이나 2005년이었던 웹사이트들은 2011년 조사 당시 이미 5-6년 정도 보존되고 있었던 것들이다. 이들이 2014년 현재 절반 가까이 아카이빙 대상에서 사라진 이유는 무엇인가? 처음부터 보존 가치가 없는 자료들을 그때까지 5-6년이나 보존하고 있었던 것인가? 아니면 최초 아카이빙 당시에는 중요하다고 평가했으나 이후 그렇지 않다고 하여 폐기한 것인가? 실제로 2011년과 2014년 사이 OASIS의 웹 자원 선정기준이나 정책이 바뀌었다는 증거는 없다.

OASIS가 채택하고 있는 웹 크롤러 등을 이용한 기계적 수집에서 매번 모든 수집 웹사이트의 품질이 균일하고 우수한 것임을 보장하기 어려울 가능성이 높다. 전문가가 정기적으로 검증하지 않는다면 누적되는 웹 자원들의 진정한 가치를 확보하지 못할 수도 있다. 앞 절에서 수집 웹사이트의 수량적 성장과 관련하여 언급한 것처럼 이전에 수집된 웹사이트들의 품질이 적절하지 않으면 제적하는 것이 맞다. 그리고 그럴 경우 삭제되는 웹사이트에 대한 정보 및 통계를 제공하는 것이 바람직하다.

5. 수집 웹사이트의 최신성

〈표 4〉 '최신 수집자료' 웹사이트 리스트

	2014년 1월 22일	2014년 11월 10일
1	제자들 국제학교	DC2009-SEOUL
2	한국교육과정평가원 대학수학능력시험	경산시
3	Jiri Lim	경북대학교 IT대학 컴퓨터학부
4	레크맨	KNU 총학생회
5	영등포 화교소학	(경북대학교 상주캠퍼스) 건설환경 공학전공
6	미대입시사	경북대학교 사회과학대학
7	전남역사교사 모임	경북대학교 법학대학 법학부
8	Daegu International School	KNU 미술학과
9	영어나라 아라모드	경북대학교 인문대학 독어독문학과
10	Gyeongnam International Foreign School	경북대학교 유럽어교육학부 독어교육전공
11	CU Info 사이버대학 종합정보	경북대학교 농업생명과학대학
12	Rainbow International School	경기대학교 평생교육원(수원)

OASIS 수집 웹 자원의 최신성을 확인하기 위하여 2014년 1월 22일, 4월 12일, 11월 10일 세 차례에 걸쳐 OASIS '최신 수집자료' 웹사이트를 검토하여, 공개된 20건 중 처음 12건

으로 <표 4>를 구성하였다. OASIS '최신 수집자료'는 우선 4건의 웹사이트 스냅 샷을 직접 디스플레이 하며, 화살표를 클릭함으로써 화면을 좌우로 이동시킬 수 있게 한다. 각 웹사이트 화면 아래 사이트명을 클릭 하면 현재 그 사이트로 연결되고, 스냅 샷을 클릭 하면, 아카이빙 시점의 사이트 메인 화면을 갈무리하여 보여주는 썸네일 스냅 샷, URL(링크), 주제 분류, 초록 정보를 포함하는 레코드 '상세정보'를 볼 수 있다. 여기에서 URL은 활성화 되어 있어 현재 해당 웹사이트로 바로 갈 수 있게 한다. 각 레코드 아래쪽에는 '웹사이트 아카이빙' 관련 정보가 주어지며, 아카이빙 버전, 날짜 및 '수집 메인 화면'으로 갈 수 있는 링크가 제공된다.

'최신 수집자료' 웹사이트에서는 몇 가지 문제점이 관찰되었다. 첫째, 아카이빙 최신성의 정의이다. <표 4> 왼쪽 칼럼의 최초 웹사이트 4건의 아카이빙 날짜는 모두 2013년 10월 30일이다. <표 4>에는 포함되지 않았으나 4월 12일 디스플레이 웹사이트 중 1건은 2013년 12월 2일, 다른 3건은 모두 2013년 11월 6일에 아카이빙 되었다. 이 웹사이트들의 실제 아카이빙 시점과 OASIS 상 공개 시점은 적어도 3-5개월 정도 간격이 있었다. 이들을 다시 11월 10일 관찰했을 때, 이전 4월에 디스플레이 되었던 웹사이트 3건은 동일하였고 '경산시' 웹사이트만 추가되었다. 얼마나 자주 OASIS가 최신 웹사이트를 수집하는지 알 수 없으나, 4월과 11월에 똑같은 웹사이트들이 '최신 수집자료'로 디스플레이 된 것은 최신성의 정의에 의문을 갖게 한다.

두 번째 문제점은 이들의 선정 이유이다. 이들이 과연 OASIS가 천명한 목표에 부합하는 "가치 있는 인터넷 자료"이며, "국가적 차원에서 수집 · 축적하여 미래 세대에 연구 자료로 제공"(국립중앙도서관. OASIS 2009) 될 만한 자료인가 하는 의문이다. <표 4> 왼쪽 칼럼에 보이는 2014년 1월 웹사이트 중 레코드 1은 '제자들 국제학교'로서 '상세정보'에 따르면, "...성경말씀을 토대로 한 기독교 신앙 안에서... 세계적인 지도자를 양육함을 목적으로 설립"된 학교이다. 레코드 2는 '한국교육과정평가원 대학수학능력시험' 웹사이트이며, 레코드 3은 'Jiri Lim'으로 2010년 3월 방송통신심의위원회 청소년 권장 사이트로 지정된 적이 있는 지리교사의 개인 홈페이지이다. 레코드 4는 '레크맨'으로서 "...초 · 중 · 고등학교 선생님들과 전문적으로 레크레이션 활동을 하시는 분들을 위해 만들어진 레크레이션 전문자료 사이트"이다. 나머지에는 상업 사이트임이 명백한 것들이 몇 개 포함되어 있으며, 국제학교나 외국인학교, 사이버학교 등의 웹사이트가 포함되어 있다.

<표 4> 오른쪽 칼럼의 2014년 11월 10일에 관찰한 12건의 웹사이트들 중 레코드 1, 2, 12를 제외하고는 모두 경북대학교 대학이나 학과 홈페이지이다. <표 4>에는 넣지 않았으나 공개된 20건 가운데 나머지 8건에는 'koreastory.com', '지리세계', '독서교재' 등 상업적 웹사이트들이 들어 있다. 상업적 사이트에도 국가적 가치 있는 자료가 될 만한 것은 있겠지만, 이들의 진정한 가치와 선정기준 적합성은 확실하지 않다.

세 번째 문제점은 웹사이트 공개와 관리에 관한 것이다. 2014년 4월과 11월 두 번 다 '최신 수집자료' 레코드 1로 디스플레이 된 'DC2009-SEOUL'(http://dc2009.kr)은 아카이빙 '수집 메인 화면'으로 가면 2013년 12월 2일 아카이빙 버전 1 '2009 International Conference on Dublin Core and Metadata Applications, 12-16 October 2009, SEOUL, KOREA' 사이트의 'DC-2009 "Semantic Interoperability of Linked Data"' 페이지를 볼 수 있다. 그러나 '상세정보' 안의 URL 링크를 따라가 보면 '앞니 레미네이트 강남' '부산아파트 담보대출' 등 리스트가 나타나고, 화면 아래 URL 링크도 마찬가지다. OASIS의 모든 아카이빙 된 레코드 화면 왼쪽 상단에는 '이 사이트는 과거 시점의 보존 파일로 외부 링크, 검색창 등이 일부 작동하지 않을 수 있습니다'란 메시지가 나온다. 실제로 OASIS가 특정한 웹사이트를 수집하여 아카이빙 이후에 해당 웹사이트 변화 혹은 해당 URL의 용도 변경에 책임을 질 수도 없고 그럴 필요도 없을 것이다. 그러나 공식적으로 수집하여 링크까지 걸어놓은 웹사이트가 'DC2009-SEOUL'과 같은 내용을 보여준다면, 이처럼 '면책'을 선언하는 메시지가 있다 하더라도 OASIS의 공신력 자체에 아무런 영향이 없다고는 하기 어렵다.

아울러 'DC2009-SEOUL'이라는 특정한 웹사이트에만 해당될 수도 있겠으나, 웹사이트 수집 시점과 관련된 추가적 의문점을 가질 수 있다. 2009년 10월에 열렸던 국제 컨퍼런스의 공식 웹사이트가 어째서 4년이나 지난 2013년 12월 2일에야 비로소 수집되고 최초 아카이빙 되었는가 하는 점이다. 이 국제 컨퍼런스가 더블린코어 메타데이터와 관련된 중요한 국제 행사로서 처음으로 우리나라에서 열렸고, 더욱이 OASIS 운영 기관인 국립중앙도서관이 후원 기관이었음에도 아카이빙 시간차가 발생한 이유는 무엇인가? 또한 이 컨퍼런스 웹사이트가 2013년 12월 2일 최초 아카이빙 되었다면, 그 이후 해당 URL의 소유권 변화가 생겼고, 부적합한 콘텐츠로 연결된다는 의미일 것이다. 그렇다면 OASIS는 아무 상관이 없는가 하는 점이다. 특정 시점의 웹사이트만을 보존하는 것이고 이후에는 그 웹사이트와 관련된 콘텐츠에 책임을 지지 않을 것이라면, OASIS 상에서 URL 링크를 활성화 시키지 않는 것이 옳다.

상기한 의문점들은 결국 OASIS 웹사이트의 수집 기준과 절차, 관리에 대한 문제로 귀결될 수밖에 없다. OASIS가 도대체 어떤 웹사이트를 언제, 어떤 방법으로 수집하고, 어떻게 유지 관리하는가에 대한 총체적인 점검이 필요하다.

6. 수집 웹사이트 통계의 정확성

OASIS 수집 웹사이트 통계가 정확하지 않다는 것은 이전 연구(윤정옥 2011)에서도 지적되었다. 기본적으로 한 개의 아카이빙 된 웹사이트는 복수의 인스턴스를 가질 수 있다. 인스턴스는 변화하는 콘텐츠를 반영하기 위해 계속해서 캡처되는 아카이빙 스냅 샷을 의미한다

(UK Archive 2014b). 호주의 PANDORA나 영국의 UK Web Archive 등은 아카이브와 인스턴스를 명확히 구별하여 정의할 뿐 아니라 통계도 각기 제시한다. 그러나 OASIS는 개별 웹사이트와 인스턴스를 구별하지 않아 종종 중복이 나타난다.

예를 들어 2014년 4월 수집 웹사이트 리스트에서 무작위로 추출한 사이트명 '한국통합물류협회'를 검색한 결과, 동일한 URL (http://www.koila.or.kr)을 갖는 동일한 기관이 각각 2012년 1월 19일과 2012년 7월 9일에 각각 수집되었고 2건의 별개 웹사이트로 검색되었다. '나눔재단(http://www.nanu.or.kr)'도 2012년 1월 19일과 2012년 8월 3일에 각각 2건의 별개 웹사이트로 아카이빙 되었다. 이 2개의 '나눔재단' 상세정보의 '수집 메인 화면'은 동일하게 '3 captures: 2011/07/29-2012/08/3'이라는 정보를 포함하고, 3차례 캡처 가운데 2차례 캡처된 것, 즉 2개의 인스턴스를 2건의 별개 수집 웹사이트로 간주하고 있음을 확인할 수 있었다.

최초의 아카이빙 버전과 캡처 기록상 일자의 불일치 문제도 있다. 예를 들어 '유네스코한국위원회' 검색 결과는 4건으로 '1. 유네스코 한국위원회, 2. 유네스코한국위원회, 3. 유네스코한국위원회, 4. 유네스코한국위원회 청년사업'을 포함하였다. 단지 띄어쓰기 차이만으로 3건이 별개 웹사이트로 수집 및 보존되었으며, 이들의 상세정보를 보면 각각의 아카이빙 버전 일자와 캡처 일자에 차이가 났다. 레코드 1(아카이빙 버전 1: 2012-01-18)은 '3 captures 2011/06/24-2012/06/12)', 즉 3차례, 레코드 2(아카이빙 버전 1: 2011-12-24)은 '9 captures 2011/06/24-2012/09/6)', 즉 9차례, 레코드 3(아카이빙 버전 1: 2012-06-21)은 '9 captures 2011/06/24-2012/09/6)', 즉 9차례 캡처된 것으로 나타났다. 레코드 4 '유네스코한국위원회 청년사업'(아카이빙 버전 1: 2012-06-12)은 '3 captures 2011/06/24-2012/06/12'로 2011년 6월 24일부터 2012년 6월 12일 사이 모두 3차례의 캡처가 이루어졌다. 말하자면 각 웹사이트의 최초 실제 캡처는 2011년 6월이나 12월에 이루어졌고, 아카이빙 버전 1은 최초 캡처 버전이 아니라 그 후인 2011년 말부터 2012년 초반 사이 캡처된 나중 것들이다. 이들의 '수집 메인 화면'에서는 각각 아카이빙 버전 1에 명시된 시점의 웹사이트를 볼 수 있고, 이후 캡처 시점 웹사이트 스냅샷도 볼 수도 있다. 그렇다면 어째서 예를 들어 레코드 1 (아카이빙 버전 1: 2012-01-18)의 3차례 캡처에서 최초 캡처 날짜인 2011년 6월 24일이 아니라 거의 반년이나 지난 시점의 2012년 1월 18일 캡처 버전이 '아카이빙 버전 1'로 간주되는지 알 수 없다.

이러한 중복이나 불명확한 아카이빙 버전 표시의 사례가 전체 수집 웹사이트들 가운데 얼마나 나타날지 이 연구에서 확인할 수 없었으나 분명 문제가 있는 것은 사실이다. 중복이 있다면, 그 규모가 얼마나 큰지, 그것이 단순 오류인지, 의도적인 숫자 부풀리기인지는 향후 전수조사로서만 답을 얻을 수 있을 것이다.

Ⅳ. 맺음말

OASIS는 국립중앙도서관이 지난 2004년부터 운영하고 있는 국가 디지털 자원 아카이브이다. OASIS가 수집한 웹사이트들을 2013년 12월부터 2014년 11월 사이 검토한 결과, 다음과 같이 몇 가지 문제점을 관찰할 수 있었다:

첫째, 지난 10년 사이 OASIS 수집 웹사이트의 수량적 성장은 괄목할 만하지만, 콘텐츠의 품질은 의문시된다. 특히 '최신 수집자료' 및 '주제별 브라우징' 웹사이트 등은 "저작자 혹은 발행자의 권위 및 학술적 가치"와 같은 OASIS의 선정 근거에 따라 국가적 디지털 자산으로서 적합한 콘텐츠 가치를 가진 웹 자원이 수집되었는지 의문을 갖게 한다.

둘째, 2013년 12월 시점의 KDC 10개 '주제별 브라우징' 리스트는 '사회과학'이 전체의 63.6%를, 그 안에서 '정치학' 소주제가 34.7%(전체의 21.4%)를 각각 구성할 정도로 수집 웹사이트의 주제 불균형이 심각함을 보여준다. 현재는 주류 이하 소주제 리스트는 제공하지 않으며, 이들의 '테마별 브라우징' 리스트에서는 웹사이트의 절반 정도를 8개 테마로 분류하였으나, 그 중 83.2%가 '기타'로 분류되어, 수집 웹사이트의 부분적 공개, 부적합한 테마 분류 등 문제점을 나타냈다.

셋째, 수집 웹사이트의 지속성 여부도 확실하지 않았다. 2011년 공개 웹사이트들 중 사이트명이 확인 가능한 55건은 2014년 4월 현재 30건(54.5%)만이 남아있다. 수집 웹사이트의 누계는 증가하고 있지만 이전 수집 자료들 중 일부만 남아있고, OASIS 주제별 통계 수치도 증감이 있어, 수집과 제적이 동시에 진행되는 것으로 보인다. 그러나 재심이나 재평가 여부는 알려진 바 없다.

넷째, 수집 웹사이트의 최신성 정의 또한 분명하지 않았다. '최신 수집자료' 웹사이트들의 아카이빙 시점과 공개 시점 사이에 4-5개월 시간차가 있고, 2013년 수집된 웹사이트가 2014년 11월에도 최신 수집 자료로 공개되기도 하였으며, 웹사이트 자체는 2009년도 것이기도 하는 등 최신성의 범위가 모호하였다.

다섯째, OASIS 수집 웹사이트 통계의 정확성을 확인하기 어렵다. 무작위로 추출한 일부 사례에서 아카이빙 시점이 다른 동일 웹사이트가 복수의 별도 레코드로 간주되었거나, 웹사이트 상세정보에 나타난 '수집 메인 화면'의 최초 캡처 날짜와 '아카이빙 버전 1' 날짜가 일치하지 않는 경우가 확인되었다. 특정 웹사이트가 여러 차례 캡처 되거나 복수의 아카이빙 버전으로 보존되었어도 이들을 별개 웹사이트로 처리하는 등 기본적 문제가 지속되고 있다.

이러한 문제점들은 이전 연구들(김유승 2008; 윤정옥 2011; 노영희, 고영선 2012)에서도 지적된 바 있으며, 현재 상황은 연구들에서도 제안된 개선방안들이 전혀 고려되지 않았음을

입증한다. OASIS는 계속 지적된 문제점들을 시발점으로 하여 전체 콘텐츠를 체계적으로 분석 및 평가하고 수집 방법과 절차 등 현황을 철저히 검증해야 할 것이다. 구체적 개선 방안은 검증 이후 OASIS 운영 주체인 국립중앙도서관이 자체적으로 수립해야 하겠지만, 이 연구에서 제안 가능한 개선 방향은 다음과 같다:

첫째, 무엇보다 OASIS 출범 당시 천명한 수집기준을 충실하게 적용하기만 해도 질적인 성장을 보장하고, 의심스러운 콘텐츠의 수집을 최소화할 수 있을 것이다. OASIS 출범 당시 여러 연구자들(서혜란 2004; 이혜원 2005)이 디지털 문화유산의 아카이빙 정책 방향을 제시하였고, OASIS 정책은 국가 차원에서 수집할 만한 콘텐츠의 가치와 적합성을 명백히 선언하였다. 선언한 바대로, 충실한 정책 실행이야말로 가장 단순하면서도 분명한 해결책이 될 수 있을 것이다.

둘째, 수집 자원의 급속한 양적 성장에 대한 집착보다는 진정한 가치를 가진 양질의 콘텐츠에 초점을 맞춘 완만한 성장을 인정할 수 있어야 한다. OASIS의 운영주체는 매년 사업의 성과와 목표 달성 여부를 수치로 증명해야 하는 부담이 있겠지만, 현재뿐만 아니라 미래의 이용자들이 동의할 만한 질적 가치를 갖지 못하는 콘텐츠는 모아두어도 별 의미가 없다. 그다지 가치가 인정되지 않을 것들을 다량 수집하는 데 급급함으로써 오히려 가치 있는 것들이 앞서 언급한 '디지털 블랙홀'에 사라질 수도 있다는 우려를 정책에 반영하고, 질적 수집에 더욱 주목해야 할 것이다.

셋째, 수집 자원의 품질을 제고하기 위해서는 수집 방법을 점검할 필요가 있다. 매년 디지털 자원 수집과 서비스 유지관리에 할당된 한정된 예산 안에서 불가피하게 여겨진 기계적 수집이나 최저가 입찰에 의존한 외주 업무 등 절차적 문제를 재고할 필요가 있다. 노영희와 고영선(2012)의 제안대로 구체적 수집 지침을 새로이 마련하는 것도 필요하다. 무엇보다 전문가가 적극 개입하여 단순한 검수가 아니라 콘텐츠 품질을 총체적으로 검토하고, 일반적 장서 개발과 같이 제적도 적절히 시행해야 한다.

OASIS가 출범한지 10년이 되었다. 국민의 세금으로 운영되며 국가 지식자원 수집과 보존의 책임을 위탁받은 국립중앙도서관이 이처럼 국가적으로 의미 있는 서비스를 제대로 운영해 왔다는 확신이 있다면, 10주년은 기념하고 자랑할 만한 시점이다. 하지만 무엇인가 하고는 있다는 명분만 근근이 유지해오고 있었다면 철저한 점검과 반성이 필요한 시점이다. 오류와 시행착오가 있었다면 이를 객관적으로 평가하고 문자 그대로 미래를 위하여 개선방안을 모색해야 할 것이다.

참고문헌

국립중앙도서관. 『2010 국립중앙도서관연보』. 서울: 국립중앙도서관, 2011.

국립중앙도서관. 『2012 국립중앙도서관연보』. 서울: 국립중앙도서관, 2013.

국립중앙도서관. OASIS. 2009. OASIS 소개. 개요. <http://www.oasis.go.kr/intro/intro_overview.jsp> [cited 2009.5.22]

국립중앙도서관. OASIS. 2013. OASIS 소개. 자원 수집 지침. <http://www.oasis.go.kr/intro_new/intro_selguide.jsp> [cited 2013.3.22]

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

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

노영희, 고영선. 2012. OASIS의 선정지침 개선(안)에 관한 연구. 『한국비블리아학회지』, 23(93): 105-137.

서혜란. 2004. 『디지털 납본제도 방안』. 서울: 국립중앙도서관.

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

윤정옥. 2011. 웹 아카이브 OASIS의 현황에 관한 연구. 『정보관리연구』, 42(3): 95-116.

이소연. 2008. 믿을 수 있는 디지털 아카이브 인증기준: OASIS 적용사례. 『정보관리학회지』, 25(3): 5-25.

이혜원. 2005. 『온라인 디지털 자원 구축 사례: 국립중앙도서관을 중심으로』. 서울: 국립중앙도서관.

한국. 통계청. 국가통계포털. 2014. KR 도메인수(연도별) (수치). 자료 갱신: 2012.12.21. <http://kosis.kr/statisticsList/statisticsList_01List.jsp> [cited 2014.11.9].

Beagrie, Neil. *National Digital Preservation Initiatives: An Overview of Development in Australia, France, the Netherlands and the United Kingdom and of Related International Activity.* Washington, D.C.: Council on Library and Information Resources and Library of Congress, 2003. <http://www.clir.org/pubs/reports/pub116/pub116.pdf> [cited 2014.12.4].

Brazier, Caroline. 2013. Born.digital@british.library: the opportunities and challenges of implementing a digital collection development strategy. Paper presented at: *IFLA WLIC 2013-Singapore-Future Libraries: Infinite Possibilities in Session 198-National Libraries*. <http://library.ifla.org/222/1/198-brazier-en.pdf> [cited 2

014.11.4].

British Library. *From Stored Knowledge to Smart Knowledge: The British Library's Content Strategy 2013-15.* <http://www.bl.uk/aboutus/stratpolprog/contstrat/british_library_content_strategy_2013.pdf> [cited 2014.11.4].

Crook, Edgar. 2008. <http://pandora.nla.gov.au/pan/13910/20080930-1156/conferences.alia.org.au/alia2008/pdfs/124.TT.pdf> [cited 2014.11.12].

Day, Michael. 2003. *Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust.* UKOLN, University of Bath. Version 1.0. <http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf> [cited 2014.11.4].

Gomes, Daniel, Miranda, Joao, and Costa, Miguel. 2011. "A Survey on Web Archiving Initiatives." *Proceedings of the 15th International Conference on Theory Practice Digital Libraries, Rese. Adv. Technol. Digital Libraries*: 408-420. <http://sobre.arquivo.pt/sobre-o-arquivo/a-survey-on-web-archiving-initiatives> [cited 2014.11.4].

Meyer, Eric. 2012. *Researcher Engagement with Web Archives-Challenges and Opportunities.* Technical Report, Joint Information Systems Committee (JISC). <JISC-REWA_ChallengesandOpportunities_August2010.pdf> [cited 2014.11.4].

Meyer, Eric. 2011. *The State of the Art and the Future(s) of Web Archiving*. Presented at the Oxford Internet Institute, October 7, 2011. <2011-10-07_BLWebArchivesTalk.pdf> [cited 2014.11.4].

National Library of Australia. 2014a. *Pandora Overview.* <http://pandora.nla.gov.au/overview.html> [cited 2014.11.4].

National Library of Australia. 2014b. PANDORA archive size and monthly growth. Statistics as at 26 October 2014. 2014b. <http://pandora.nla.gov.au/statistics.html> [cited 2014.11.4].

National Library of Australia. Preserving Access to Digital Information (PADI). 2014. Web Archiving. <http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/92.html> [cited 2014.11.4].

Toyoda, M., and Kitsuregawa, M. 2012. "The History of Web Archiving." *Proceedings of the IEEE 100:* 1441-1443.

UK Web Archive. 2014a. UK Web Archive statistics. Statistics at 06 November, 2014.

<http://www.webarchive.org.uk/ukwa/statistics> [cited 2014.11.6].

UK Web Archive. 2014b. *What is the UK Web Archive?* <http://www.webarchive.org.uk/ukwa/info/about> [cited 2014.11.4].

UNESCO. 2004. *Charter on the Preservation of Digital Heritage.* Records of the General Conference, 32nd Session, Paris, 29 September to 17 October 2003. Paris: UNESCO, 2004. <http://unesdoc.unesco.org/images/0013/001331/133171e. pdf> [cited 2014.11.6].

국한문 참고문헌의 영문 표기
(English translation / Romanization of reference originally written in Korean)

National Library of Korea. 2011. *Annual Report 2010 National Library of Korea.* Seoul: National Library of Korea.

National Library of Korea. 2013. *Annual Report 2012 National Library of Korea.* Seoul: National Library of Korea.

National Library of Korea. OASIS. 2009. OASIS Introduction. Overview. <http://www.oasis.go.kr/intro/intro_overview.jsp> [cited 2009.5.22].

National Library of Korea. OASIS. 2013. *Selection Guidelines.* <http://www.oasis.go.kr/intro_new/intro_selguide.jsp> [cited 2013.3.22.].

Kim, You-seung. 2008. "A Study of Combined Web Archiving Policy : BnF's Three Layers Web Archiving Strategy." *Journal of the Korean Society for Library and Information Science,* 42(4): 159-179.

Kim, You-seung. 2007. "A Study of Legal Issues for Web Archiving." *Journal of the Korean Society for Library and Information Science,* 41(3): 5-24.

Noh, Younghee, & Go, Youngsun. 2012. "A Study on Improving the OASIS Selection Guidelines." *Journal of the Korean Biblia Society for Library and Information Science,* 23(3): 105-137.

Suh, Hye-Ran. 2004. *Policies of Digital Deposits.* Seoul: National Library of Korea.

Yoon, Cheong-Ok. 2010. "A Research on the OASIS, a Web Archive in Korea." *Journal of the Korean Society for Library and Information Science,* 44(2): 5-27.

Yoon, Cheong-Ok. 2011. "A Research on the OASIS, a Web Archive in Korea, Revisited." *Journal of Information Management,* 42(3): 95-116.

Lee, So-Yeon. 2008. "Trustworthy Repositories Audit Criteria: Self-Assessment of

OASIS." *Journal of the Korean Society for Information Management,* 25(3): 5-25.

Lee, Hyewon. 2004. *Development of Online Digital Resources: The Case of the National Library of Korea.* Seoul: National Library of Korea.

Korean Statistical Information Service(KOSIS). 2014. kr domain (Year) (Numbers). Upadate: 2012.12.21. <http://kosis.kr/statisticsList/statisticsList_01List.jsp> [cited 2014.11.9].