웹 아카이브 OASIS에 관한 고찰

Research on the OASIS, a Web Archive in Korea

윤 정 옥(Cheong-Ok Yoon)*

목 차

1. 머리말
  1.1 선행 연구
  1.2 연구의 목적과 방법
2. 다른 나라의 웹 아카이브
3. OASIS의 '자원 수집 지침'과 현황
  3.1 OASIS의 '자원 수집 지침'
  3.2 OASIS 수집 자원의 현황
4. OASIS 콘텐츠의 특성
  4.1 우리나라와의 관련성
  4.2 저작자의 평판
  4.3 정보의 유일성
  4.4 학술적 내용
  4.5 정보의 최신성
5. 맺음말

초 록

이 연구에서는 국립중앙도서관이 구축・운영하는 웹 아카이브인 OASIS의 특성과 문제점을 살펴보았다. OASIS 웹 사이트를 통해 제공되는 웹 자원 콘텐츠의 현황을 검토하고, 향후 필요한 개선 방안을 제안하였다. 이 연구를 위해 2009년 5월 20일부터 7월 2일 사이 OASIS 웹 사이트에서 '문학'과 '사회과학' 분야의 콘텐츠 및 이용 가능한 서비스를 검토한 결과, 수집된 콘텐츠의 주제 분포의 편중, 저작자/발행자 편향성과 권위의 근거 미약, 정보의 유일성과 최신성 결여, 웹 문서와 웹 사이트의 중복 수집, 학술적 가치의 근거 결여 등 문제점이 발견되었다. 주제전문가의 활용과 실명제, 메타데이터 요소 추가 및 기본적 목표의식의 지속적 확인 등이 필요함을 제안하였다.

ABSTRACT

The purpose of this research is to examine the characteristics and problems of OASIS, a web archive, developed and operated by the National Library of Korea, and then to propose how to improve the quality of its contents. An analysis of the contents in 'Literature' and 'Social Sciences' in the OASIS website shows problems, including low quality of some contents, lack of balance in subjects and creators/publishers, duplicate collections of web documents and web sites, lack of currency and evidence of unique or scholarly value of collected web resources, etc. Suggested are the use of 'real-name' of subject specialists to ensure their responsibilities in selection process and the addition of metadata elements to evidence the appropriate application of selection criteria.

키워드: 오아시스, 웹 아카이브, 국립중앙도서관, 디지털자원
OASIS, Web Archive, National Library of Korea, Digital Resources

---

* 청주대학교 인문대학 문헌정보학전공 교수(jade@cju.ac.kr)
논문접수일자: 2009년 7월 15일 최초심사일자: 2009년 8월 24일 게재확정일자: 2009년 10월 31일
한국문헌정보학회지, 44(2): 5-27, 2010. 〔DOI:10.4275/KSLIS.2010.44.2.005〕

www.kci.go.kr

1. 머리말

오늘날 웹이 많은 사람들에게 "가장 먼저 의지하는 정보원(the information source of first resort)"(Lyman 2002)이 되고 있다는 것은 부인하기 어렵다. 그리고 웹 상에서만 생산되고 유통되며 신속히 소멸되는 막대한 정보자원이 "문화유산의 중요한 부분"이 되고 있다는 점도 또한 널리 인정되고 있다. 따라서 여러 나라들은 웹 상에서만 이용 가능한 다양한 정보자원을 국가의 문화적 자산으로 간주하여 국가 차원에서 수집 및 조직하고, 국민들로 하여금 이용할 수 있게 하며, 지금과 미래세대를 위하여 이를 보존하려는 노력을 시작했다. 2000년대 초반 Day가 영국의 Joint Information Systems Committee를 위하여 웹 아카이브에 대한 서베이를 수행할 당시 미국, 영국, 호주, 오스트리아, 프랑스, 스웨덴, 핀란드 등이 이러한 사업을 선도하였고(Day 2003), 2003년부터는 이들 나라의 국립도서관들 및 미국의 Internet Archive를 포함한 12개 기관이 국제인터넷보존컨소시엄(International Internet Preservation Consortium. 이하 IIPC라 부름)을 결성하여 인터넷 콘텐츠의 수집, 보존 및 장기적 접근 보장을 위한 국제적 협력을 수행하고 있다. 또한 UNESCO는 디지털 형태로 변환되거나 디지털 형태로만 생산되는 문화, 과학, 및 정보유산을 보존하고 국제적 접근을 보장하기 위한 지침을 마련하였다(UNESCO 2003).

우리나라 국립중앙도서관의 OASIS(Online Archiving & Searching Internet Sources)는 이러한 맥락에서 "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 2009a). 국립중앙도서관은 OASIS 프로젝트를 위하여 2001년 11월 온라인 저작물 수집・보존 시스템 ISP 및 시범시스템 개발을 시작으로 하여, 2005년 12월 1일 웹기반 OASIS 시스템 확장 및 개선 사업을 완료하였고, 2006년 2월 1일 OASIS 홈페이지 대국민 서비스를 실시하기 시작하였다. OASIS는 2006년 6월 문화관광부 "5대 혁신브랜드"로 선정되었고, 2006년 12월 업무평가 시 2006년분 웹 자원 63,270건 수집을 보고하였으며, 2007년 12월 업무평가 시 2007년분 웹 자원 100,512건 수집을 보고한 바 있다(국립중앙도서관 2009b). 특히 2006년 사업보고서는 "3-1 성과목표의 달성도"에서 2005년 수집실적이 목표량의 137%를 달성하였다고 하여, 처음부터 사업이 성공적으로 수행되고 있다는 점을 강조하였다. 국립중앙도서관이 2009년 5월 29일 현재 OASIS 웹 사이트[1]를 통해 공개하고 있는 웹 자원의 총수는 웹 문서 48,547건과 웹 사이트 257종으로서, 이전에 수집하였다고 한 웹 자원의 전체 규모보다는 훨씬 적기는 하지만, 일단은 누구든 웹 상에서 이용할 수 있게 한다는 점에서 의의가 있다고 하겠다.

그러나 OASIS가 본격적으로 웹 자원을 수집하고, 인터넷 웹 사이트를 통해 콘텐츠를 공개하기 시작한 이래, 실제로 OASIS에 수집된

1) 〈www.oasis.go.kr〉.

www.kci.go.kr

콘텐츠의 내용과 품질이 국가 디지털자원의 보존 대상으로서 과연 적정한 수준을 달성하고 있는지, 또한 이용자를 위한 시스템과 서비스가 적절히 기능하고 있는지에 대해서는 그동안 논의된 적이 없다. 매년 국립중앙도서관의 사업성과에 대한 내부적 평가가 있었겠지만, 이용자 측면에서 외부적 검토가 필요한 시점이라고 하겠다.

1.1 선행 연구

OASIS 프로젝트를 시작할 당시 국가적 웹 아카이브의 필요성이나 웹 아카이빙의 정책적 목표나 전략 개발 등 측면에 대한 연구들이 수행되었으나(서혜란 2004; 최원태 2005; 이혜원 2004 등), OASIS의 본격 출범 이후에는 그와 관련된 학술적 연구의 수는 손꼽을 정도이다.

최근 이소연(2008)은 미국의 RLG와 국립기록관리관(NARA)이 제시한 디지털 아카이브의 평가도구인 "TRAC(Trustworthy Repositories Audit & Certification)"을 사용하여 OASIS를 검토한 바 있다. 이 연구는 국제적으로 인정된 평가도구의 기준에 따라 OASIS가 디지털 아카이브로서 웹 자원의 수집, 보존, 관리 등 전 단계에서 내부적으로 갖추어야 할 조건들을 분석한 것이다. 연구자는 OASIS의 조직 인프라(24개 영역), 디지털 객체관리(44개 영역), 기술 인프라(16개 영역)는 그 조직 및 기술 인프라 측면에 비하여 웹 자원의 장기 보존 기능을 포함한 객체 관리 인프라가 취약한 상태임을 지적하였다.

김유승(2007)은 분산성, 휘발성, 역동성, 심층 웹 및 진화하는 웹이라는 웹 정보자원의 기본적 특성을 논하는 가운데, OASIS가 선택적으로 웹 자원을 수집하는 방법의 문제점을 지적하였고, 수집과 보존이라는 일차적 목적을 넘어서서 연구 활동 및 다양한 서비스와 활용에 대해 고민해야 한다고 강조하였다.

김유승(2008)은 웹 아카이빙의 유형을 선택적, 복합적 및 전체 도메인 아카이빙의 세 개로 구분하고, 복합적 아카이빙을 채택한 대표적 기관으로서 프랑스 국립도서관의 사례를 검토하고, 더불어 OASIS에 대한 '비판적 분석'을 수행하였다. OASIS의 수집 범위, 방법, 품질의 측면에서 주제별 정보자원에 대한 수집이 적극적이지 못하며, 제한적이고 소극적인 수집에 머물고 있다는 점, 단일 웹 문서와 사이트의 정보량 차이가 크며, 수집된 웹 사이트가 만족스럽지 못한 수평적, 수직적 품질을 보인다는 점 등을 지적하였다. 연구자는 또한 OASIS가 국가 대표 아카이브로서 "가능한 한 가장 넓고 심도 깊은" 수집 범위 설정, 다양한 아카이빙 정책의 적용 및 타 기관과의 협력적 웹 아카이빙 전략 수립 등이 필요함을 제안하였다.

1.2 연구의 목적과 방법

이 연구는 OASIS 웹 사이트를 통해 제공되는 웹 자원 콘텐츠와 서비스의 현황과 특성을 살펴보고, 향후 국가 자산으로서 가치 있는 디지털 자원의 수집과 제공을 위해 필요한 개선 및 발전 방안을 제안하고자 하는 목적으로 수행되었다.

이 연구를 위해 먼저 2009년 5월 20일부터 7월 2일 사이 OASIS 웹 사이트에서 '문학'과 '사회과학' 분야의 웹 자원 콘텐츠에 관련된 특성을 검토하였고, 이를 10월 초에 다시 검토하

www.kci.go.kr

여 변동 여부를 확인하였다.

이 연구는 원래 연구자가 OASIS 웹 사이트를 개인적으로 이용하는 과정에서 관찰된 문제점을 단계별로 정리하는 데서 시작되었다. 다시 말하면 처음부터 OASIS 웹사이트에서 제공하는 콘텐츠와 서비스를 평가하려는 목적을 갖고 구체적이며 객관적인 웹 아카이브의 평가기준을 적용한 것이 아니다. 따라서 여기에서는 OASIS의 '자원 수집 기준' 항목을 적용하여 그동안 관찰한 콘텐츠의 특성을 정리하는 데 초점을 맞추었다. 또한 이러한 연구 자체의 제한점을 감안하여 이후 OASIS 운영 주체인 국립중앙도서관이나 관련 기관에서 보다 체계적인 평가나 분석을 시행하는 데 도움이 될 수 있도록 현재 시점에서의 문제점을 기술하고 의견을 제시하였다.

2. 다른 나라의 웹 아카이브

다른 나라에서 구축된 웹 아카이브들의 특성에 대해서는 최근 일본국립국회도서관 관서관(關西館) 전자도서관과의 다케다 가즈야(武田和也 2008)가 정리한 자료가 있다. 다케다는 이들을 반드시 "국가대표"라고 명시하지는 않았지만, 한국을 비롯하여 모두 열여덟 개 나라에서 대부분 국립도서관 혹은 국가대표도서관이 유지하는 것들을 살펴보았다. 웹 자원의 수집 범위는 웹 아카이브마다 다소 차이가 있어, 이들 가운데 싱가폴, 프랑스, 노르웨이, 스웨덴, 오스트리아는 웹 자원을 포괄적으로 수집한다고 하였으나, 미국, 캐나다, 호주, 영국 등 나머지 국가들은 특정한 범주나 유형의 웹 자원만을 선택적으로 수집한다고 하였다. 한편 웹 아카이브의 자원 수집과 유지관리 주체도 상이한 점이 있다. 예를 들어 중국이나 프랑스처럼 한 개의 도서관이 웹 자원 수집을 전담하고 있는 경우도 있지만, 미국이나 대만처럼 복수의 기관들이 각기 웹 아카이브를 유지하는 경우도 있고, 호주나 캐나다처럼 하나의 국가적 웹 아카이브를 위한 자원 수집에 복수의 도서관과 문서관 등이 함께 참여하는 경우도 있었다.

이들 웹 아카이브 가운데 몇 개가 가진 목표와 자원 수집의 기준을 각각 살펴보면, 먼저 일본의 웹 아카이브는 국립국회도서관이 주도한 '인터넷자원선택축적실험사업(インターネット資源選擇蓄積實驗事業)'이라는 명칭의 실험 프로젝트로서 2002년 4월 시작되었고, 2006년 7월 WARP(Web Archiving Project)로서 본격적으로 사업화되었다. WARP는 "인터넷상에 공개되어 있는 수많은 유용한 정보자원을 문화자원으로서 장래 세대를 위하여 보존하는 프로젝트"임을 표방하고 있다. WARP는 국가기관, 도・도・부・현(都道府縣), 정령지정도시(政令指定都市), 시・정・촌합병(市町村合併), 법인・기구, 대학 및 이벤트의 홈페이지와 전자잡지를 수집하고 있다. 2008년 6월 현재 국가기관 홈페이지 51종, 도・도・부・현 홈페이지 34종, 정령지정도시 14종, 시・정・촌합병 1,714종, 법인・기구 176종, 대학 142종, 이벤트 532종, 전자잡지 1,557종 등 모두 3,741종을 수집하였다(日本. 國立國會図書館 2009a).

호주의 웹 아카이브인 PANDORA는 1996년 시범사업으로 호주국립도서관(National Library of Australia)이 처음 구축하기 시작하였다. 그러나 현재는 State Library of Queensland, State

www.kci.go.kr

Library of South Australia 등과 같은 주립도서관 및 National Film and Sound Archive, Australian War Memorial 등과 같은 문화수집기관을 포함한 아홉 개의 기관들이 협력하여 구축하고 있다. PANDORA 아카이브의 목표는 다음과 같다(National Library of Australia and Partners 2009a):

> "호주에 관한, 사회적, 정치적, 문화적, 종교적, 과학적 혹은 경제적 중요성 및 호주에 대한 관련성을 가진 주제에 대하여 호주인 저자에 의한, 혹은 인정된 권위를 가진 호주인 저자에 의한 것이며 국제적 지식에 기여하는 선별된 온라인 간행물과 웹 사이트를 수집하고 장기적 접근을 제공한다."

상기한 목표를 가진 PANDORA는 "선택적 아카이브(a selective archive)"로서 호주의 모든 온라인 간행물과 웹 사이트를 수집하려고 시도하지는 않으며, 오직 "중요성이 있으며, 장기적인 연구(research) 가치가 있는 것"만을 수집한다고 하였다(National Library of Australia and Partners 2009c). PANDORA 아카이브는 2009년 4월 27일 현재 아카이브된 타이틀 21,862건, 아카이브된 인스턴스[2] 45,181건, 파일 66,545,542개를 소장하고 있고, 소장 파일의 규모는 2.88TB에 이른다(National Library of Australia and Partners 2009b).

미국의회도서관 웹 아카이브(The Library of Congress Web Archives. LCWA)는 초기 개발 프로젝트 시 MINERVA라고 불렸다. 이 웹 아카이브는 "지정된 주제에 대한 웹 기반 정보를 대표하기 위하여 주제전문가가 선정한 아카이브 된 웹 사이트의 컬렉션"으로서 구축되었다. LCWA는 "다음 세대의 연구자를 위해서 디지털 자원을 평가, 선정, 수집, 목록, 접근 제공 및 보존하기 위한 미국의회도서관의 지속적 노력의 일부"임을 천명하고 있다. LCWA는 MINERVA로서 맨 처음 구축한 '2001년 9.11 사건(September 11, 2001) 웹 아카이브'를 포함하여, 2009년 6월 현재 '수단 다르프 위기 웹 아카이브(Crisis in Darfur, Sudan, Web Archive, 2006)', '이라크 전쟁 아카이브 웹 아카이브(Iraq War, 2003 Web Archive)', 2000년, 2002년, 2004년 미국 선거 웹 아카이브 등 모두 12개 아카이브를 포함하고 있다. 이 가운데 예를 들면, '수단 다르프 위기 웹 아카이브'는 수집기간은 2006년 3월 30일부터 같은 해 11월 20일까지로, "About SUDIA - Sudanese Development Initiative", "ADL Curriculum Quarterly: Spring 2005. Lesson 4: Genocide in Darfur: Is the World Doing Enough?", "AfricaFocus home page" 등 수록된 사이트 수는 216개에 이른다(Library of Congress Web Archives 2009).

이들을 비롯한 여러 나라의 웹 아카이브에서 어떤 수준의 디지털자원을 어떤 방법으로 수집하고, 어떤 방식으로 제공하는가는 차이점을 보이는데, 이는 2008년 IIPC의 연구에서 나타난 바 있다. IIPC는 최초 회원은 12개국 기관이었지만, 이후 웹 아카이빙을 수행하는 기타 도서

2) 아카이브된 인스턴스(archived instance)는 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. 각각의 새로운 카피는 '아카이브된 인스턴스'라고 불린다.

www.kci.go.kr

관, 박물관, 아카이브, 문화유산기관 등으로 회원을 확대하였고, 2008년 3월 이들의 프로파일을 조사하였다. 이 조사에는 2008년 현재 39개 회원 기관 중 35개 기관이 응답하였고, 이들 가운데 국립도서관이 24개, 국립 아카이브가 1개 있었고, 그밖에는 다른 유형의 도서관, 대학, 박물관, 서비스 제공자, 기타 기관 등이 포함되었다. 이들의 웹 아카이빙 범위는 광범한 혹은 "대량 크로울링(bulk crawling)"(20.0%, 7개), 전체 국가 도메인(40.0%, 14개), 지역 도메인(17.1%, 6개), 주제/사건/선별 아카이빙(94.3%, 33개), 기타(42/9%, 15개) 등으로 다양하였다. 또한 간행된 선정 지침이나 수집 개발 지침을 가진 기관은 26.5%(9개)였고, 지침이 있지만 웹 사이트에 공개하지 않은 기관이 23.5%(8개), 지침이 없는 기관이 29.4%(10개), 지침이 없지만 내부적 용도로만 있는 기관이 20.6%(7개)였다(Grotke 2008, 8). 또한 수집한 웹 자원에 대한 접근 제공 여부 및 수준도 다양하였다. 모든 것을 "비공개 아카이브(dark archive)"에 소장하고, 수집기관에서 아카이빙 및 기타 직원들만 콘텐츠 접근(16.1%, 5개), 완전한 비공개 아카이브로 접근 불허(3.2%, 1개), 수집기관 내에서 연구자들에게만 접근 개방(22.6%, 7개), 관내에서 연구자 및(인터넷을 통해) 관외에서 누구에게든 개방(29.0%, 9개) 및 기타(67.7%, 21개) 등이었다.

이처럼 나라마다 수집 대상으로 하는 디지털 자원의 종류와 범위가 다르고, 수집하는 방법이나 접근 허용 수준도 다르지만, 이들이 공통적으로 지향하는 바는 동일하다고 하겠다. 말하자면, 자국에서 생산되고 공개된 디지털자원은 적절히 수집하고 보존해서 다음 세대로 하여금 접근할 수 있게 해야 하는 국가적 자산임을 인정한다는 것이다. 이를 위해서 대체로 국가대표도서관이 그 수집, 보존 및 접근 제공의 책임을 맡고 있다. 이러한 책임 의식은 미국의 회도서관이 "교육과 연구를 촉진하기 위하여 역사적 중요성을 갖는 컬렉션 자료를 수집하고, 목록하고, 보존하며 의회와 미국국민에게 제공하는 전통적인 기능은 웹 사이트를 포함한 디지털자료로 확장된다."(Library of Congress 2009)고 한 데서 간명하게 알 수 있으며, 우리나라의 OASIS가 지향하는 바도 크게 다르지 않다고 하겠다.

3. OASIS의 '자원 수집 지침'과 현황

3.1 OASIS의 '자원 수집 지침'

OASIS는 수집, 보존 및 서비스하게 되는 대상을 하나 또는 몇 개의 파일 단위로서 정보의 가치를 가지게 되는 '개별자원'과 하나의 사이트 단위로 해당 사이트의 모든 자원을 가져와 미러링하게 되는 '웹 사이트'로 구분하고, '자원 수집 지침'을 공개하고 있다(국립중앙도서관 OASIS 2009b). 김유승은 OASIS의 '자원 수집 지침'이 수집 대상 정보자원의 기준으로서 '정보 자체의 특성'과 '우리나라와의 관련성'이라는 두 가지 측면을 제시하고 있다고 지적한 바 있다(2008, 171). 이 연구에서는 김유승의 구분 방법을 차용하여 〈표 1〉과 같이 '자원 수집 지침'의 내용을 구분하고 항목화 하였다. 한편 〈표 2〉는 OASIS가 우선적으로 수집하는 자료 및 제외하는 자료를 예시한 것이다. 2008년 〈국립중앙도

www.kci.go.kr

〈표 1〉 OASIS의 '자원 수집 지침'

정보 자체의 특성	우리나라와의 관련성
현재/미래 정보요구에 대한 유용성	한국과 관련
저작자의 평판	사회, 정치, 문화, 종교, 과학, 경제적 중요성
제공된 정보의 유일성	한국의 저자가 작성
학술적 내용	- 권위 있는 한국 사람
정보의 최신성	국내외적으로 해당 학문 분야에 기여
업그레이드 빈도	
접근의 용이성	

〈표 2〉 OASIS의 수집 우선 및 제외 자료

수집 우선 자료	수집 제외 자료
중앙정부가 생산한 온라인 디지털자원	채팅사이트
대학간행물	언론사이트
회의자료, 전자저널	게시판과 뉴스그룹
기증/추천된 온라인 디지털자원	기타 기술적으로 수집이 불가능하며 수집/보존의 가치가 없는 디지털자원
최근 이슈가 되는 온라인 디지털자원	
국내 웹사이트	

서관연보〉는 실제로 OASIS를 위해 그동안 "연구보고서, 간행물, 정책자료, 통계자료 등 주로 웹사이트 내 공개자료실에서 제공되는 웹 문서와 신행정수도, 선거사이트 등 모든 주제의 웹사이트"를 수집하였다고 보고했다(국립중앙도서관 2009).

3.2 OASIS 수집 자원의 현황

OASIS 웹 사이트는 〈그림 1〉의 초기화면에서 보는 바와 같이 자료검색 및 사이트검색 창, '주제별 브라우징' 리스트, '최신자료' 및 '많이 본 자료'를 주요한 구성요소로 갖고 있다. '주제별 브라우징' 리스트는 OASIS가 현재 수집하여 보유하고 있는 웹 문서의 현황을 KDC 분류체계에 근거한 열 개 주제로 구분하여 보여주고 있다. 〈표 3〉은 OASIS가 2009년 5월 29일과 10월 9일 현재 수집 자원의 현황을 각각 보여준다. 각각의 수치는 OASIS 웹사이트 초기 화면 '주제별 브라우징' 리스트 및 다음 단계 '자료검색결과'의 '주제별 디렉토리'에서 가져온 것으로, 다섯 번째와 여섯 번째 컬럼의 화살표는 수집 웹 문서 건수 혹은 웹 사이트 종수의 증감을 의미한다.

처음 이 연구를 위한 자료 조사를 진행한 2009년 5월 당시 초기 화면에 명시된 웹 문서의 건수는 모두 48,747건이었다. 웹 문서 가운데는 기술과학 주제가 32,355건으로 가장 많았고, 그 다음은 사회과학의 10,519건이었다. 그러나 예술(1,758건)과 역사(1,127건) 분야를 제외한 다른 주제의 웹 문서는 많지 않으며, 특히 철학 분야 웹 문서는 가장 적어 단 44건만 수집되어

www.kci.go.kr

〈그림 1〉 OASIS 웹 사이트 초기화면 [접근: 2009. 6. 5. 오후 5:09]

〈표 3〉 OASIS 수집 자원의 현황

주제	2009년 5월 29일		2009년 10월 9일		
	초기 화면	주제별 디렉토리	초기 화면	주제별 디렉토리	
	웹 문서	웹사이트	웹 문서	웹 문서	웹사이트
총류	416	31	423	416	31
철학	44	3	46	44	3
종교	467	20	468	467	20
사회과학	10,519	76	10,685	(↑) 10,828	(↑) 78
순수과학	705	6	715	705	6
기술과학	32,555	80	32,675	(↑) 32,933	80
예술	1,757	15	1,764	(↑) 1,760	15
언어	521	5	523	(↓) 521	5
문학	636	11	642	636	11
역사	1,127	10	1,135	(↓) 1,131	10
합계	48,747건	257종	49,076건	49,441건	259종

있었다. 이때 '자료검색결과'의 '주제별 디렉토리'에서 '웹 문서' 항목의 수치를 별도로 확인하지 않았다. 따라서 초기 화면의 '주제별 브라우징' 리스트와 '주제별 디렉토리'에 각각 제시된 웹 문서 수집 건수가 일치하는지는 알 수 없다.

수집된 웹 사이트의 종수는 '자료검색결과'

www.kci.go.kr

의 '주제별 디렉토리'에서 '웹 사이트' 항목을 클릭함으로써 확인할 수 있었다. 5월 29일 당시 수집된 웹 사이트는 모두 257종이었다. 열 개 주제 가운데 기술과학 주제가 80종, 사회과학 주제가 76종으로 가장 많았고, 철학 주제가 3종, 언어 주제가 5종, 순수과학 주제가 6종만이 수집되어 있었다.

2009년 10월 9일 다시 OASIS 웹사이트를 확인한 결과는 초기 화면 '주제별 브라우징' 리스트의 웹 문서 건수가 모든 주제 분야에서 약간씩 증가하였다. 4개월여 사이에 총류는 7건, 철학은 2건, 종교는 1건, 사회과학은 166건, 순수과학은 10건, 기술과학은 220건, 예술은 7건, 언어는 2건, 문학은 6건, 그리고 역사는 8건이 증가하였다. 그러나 '자료검색결과'의 '주제별 디렉토리'에서 제시된 웹 문서의 건수에는 그 변동 사항이 정확하게 반영되지 않고 있다. 예를 들어 〈표 3〉에서 보는 바와 같이 초기 화면에서는 총류의 웹 문서가 423건 수집되어있다고 하였으나, '주제별 디렉토리'에는 이전과 마찬가지로 416건으로 되어 있다. 사회과학 분야는 초기 화면에서 10,685건이 수집되어 있다고 하였으나, '주제별 디렉토리'에는 10,828건으로, 기술과학 분야는 초기 화면에서는 32,675건으로, '주제별 디렉토리'에는 32,933건으로 제시되어 있다. 언어 분야는 초기 화면에는 523건이라고 제시되었으나, '주제별 디렉토리'에서는 2건 감소한 521건이라고 제시되어 있다. 역사 분야 또한 '주제별 디렉토리'에는 4건 감소한 수치가 제시되고 있다. 초기 화면의 '주제별 브라우징' 리스트와 '주제별 디렉토리'에 제시된 웹 문서의 총수는 각각 49,076건과 49,441건으로 일치하지 않는다.

한편 '자료검색결과'의 '주제별 디렉토리'에서 확인한 웹 사이트의 전체 수집 종수는 사회과학 분야를 제외하고는 아홉 개 주제 분야 전체에서 5월 29일과 10월 9일 사이에 아무런 변동이 없었다. 사회과학 분야의 웹 사이트 종수는 5월 현재 76종이었으나 10월에는 2종 증가한 78종이 제시되어 있다.

4. OASIS 콘텐츠의 특성

OASIS와 같은 국가적 차원의 웹 아카이브는 웹 자원을 수집하고 보존하는 것만이 아니라, 지금 세대와 미래 세대가 접근하고 이용할 수 있게 한다는 목표를 가지고 있다. 따라서 OASIS에 수집된 콘텐츠는 '자원 수집 지침'에서 명시된 바와 같이 현재와 미래의 정보요구에 대한 유용성을 가지는 것이라고 생각할 수 있다. 여기에서는 OASIS 웹 사이트를 통해서 검색 및 접근할 수 있는 이러한 콘텐츠의 특성을 주로 '문학' 및 '사회과학' 분야를 중심으로 살펴보도록 한다. 그 내용은 OASIS의 '자원 수집 지침'에서 제시된 '우리나라와의 관련성' 그리고 '정보자체의 특성' 관련 항목인 저작자의 평판, 제공된 정보의 유일성, 학술적 내용, 정보의 최신성 등을 중심으로 서술하였다.

4.1 우리나라와의 관련성

OASIS 콘텐츠가 '우리나라와의 관련성'을 가지는 것임은 당연하다. OASIS '자원 수집 지침'은 수집대상 디지털자원이 국가적 보존자원으로 선정되기 위해서는 "한국과 관련된 사회,

www.kci.go.kr

정치, 문화, 종교, 과학 또는 경제적으로 중요한 주제이고, 한국의 저자에 의해 쓰여진 것"이어야 하며 "또한 국내의 저명한 대학 교수나 대학 연구원 등 권위 있는 한국 사람에 의해 쓰여지거나 국내외적으로 해당 학문분야에 기여를 한 것"이어야 한다고 명시하고 있다. 이 지침에 따르면 특정한 웹 자원이 이러한 조건을 모두 만족해야만 비로소 선정되어 보존될 수 있다. 실제로도 OASIS 콘텐츠가 '우리나라와의 관련성'을 가진다는 것은 현재까지 수집된 웹 자원의 주제 분포에서 알 수 있고, 특히 그러한 특성은 '문학' 분야에서 두드러진다.

'문학' 분야를 보면 한국 관련 주제가 압도적이다. 2009년 5월 현재 모두 636건의 웹 문서가 수집되어 있는 '문학'은 KDC 분류 체계에 따라 문학, 한국문학, 중국문학, 일본문학, 영미문학, 독일문학, 프랑스문학, 스페인문학, 이탈리아문학, 기타 제문학의 10개 하부주제로 세분되어 있다. 그러나 다음 〈표 4〉에서 보는 바와 같이 이들 가운데 문학(23건), 한국문학(588건), 일본문학(4건), 영미문학(10건), 프랑스문학(1건) 및 기타 제문학(10건)의 다섯 개 주제에만 웹 문서가 수집되어 있고, 나머지 주제의 수집 웹 문서는 0건이다. 그 가운데, 한국문학이 588건(92.45%)

〈표 4〉 '문학' 분야 웹 문서의 발행자(저작자) 분석 [접근: 2009. 5. 27]

주제	총문서	발행자(저작자)	문서	수집일
문학	23	마광수	1	2006. 3. 16
		영미문학연구회	1	2006. 2. 20
		한국스피치커뮤니케이션학회	20	2005. 8. 5
		국립국어원	1	2004. 6. 4
한국문학	588	우리말배움터	1	2007. 3. 22
		동서대학교일본연구센터	1	2006. 3. 30
		사회평론사(마광수)	3	2006. 3. 15
		청하(마광수)	6	2006. 3. 15
		해냄출판사(마광수)	3	2006. 3. 15
		조선일보(마광수)	1	2006. 3. 2
		마광수	388	2006. 2-4
		에세이코리아	24	2005. 8. 26
		황국명	13	2005. 6. 7
		한국고전의세계	51	2004. 11. 9 - 18
		대한민국학술원	1	2004. 10. 5
		한국현대문학회	97	2004. 8. 6
일본문학	4	한국외대 외국학종합센터 일본연구소	1	2007. 5. 22
		동서대학교 일본연구센터	3	2006. 3. 30
영미문학	10	에세이코리아	1	2005. 8. 26
		최희섭교수의 홈페이지	8	2005. 6. 3
		대한민국학술원	1	2004. 10. 5
프랑스문학	1	한국외대아프리카연구소	1	2006. 3. 22
기타 제문학	10	한국외대아프리카연구소	10	2006. 3. 22

www.kci.go.kr

으로 절대다수를 차지하는 한편 나머지 주제의 문서들은 소수일 뿐이다. 이처럼 주제가 한국문학에 편향되어 있다는 것은 "한국과 관련된…" 혹은 "한국 저자에 의해…"라는 기준을 근거로 할 때 타당하다고 할 수 있으나, 다른 주제 분야에는 그토록 수집하고 보존할 만한 성과물이 없는지 다소 의문을 가질 수도 있다.

한편 이 연구에서 원래 분석하였던 '문학' 분야 외에 '역사', '예술' 분야 등의 한국 관련 주제 분포를 살펴보기 위해 일정기간 수시로 접근하였으나, 계속해서 OASIS 시스템 오류가 발생하여 추가 정보를 얻을 수 없었다(접근: 2009. 10. 15 - 17). 빈발하는 접근 오류 또한 중대한 문제라고 볼 수 있으나, 콘텐츠에 초점을 맞추고 있는 이 연구에서 따로 언급하지는 않기로 하였다.

4.2 저작자의 평판

OASIS에 수집된 콘텐츠의 '저작자의 평판'과 관련하여 가장 두드러지는 문제는 저작자 및 발행자의 편향성과 권위라고 할 수 있다.

먼저 저작자 및 발행자의 편향성은 '문학' 분야에서 눈에 띈다. '문학' 분야에 수집된 웹 문서 636건을 분석한 결과, 관련된 발행자는 〈표 4〉에서 보는 바와 같이 영미문학연구회, 국립국어원, 한국현대문학회, 대한민국학술원 등 모두 열다섯 개로 나타났다. 그러나 실제로 이 주제 분야는 마광수교수의 개인 웹 아카이브라고 해도 과언이 아니다. 열다섯 개의 발행자들 가운데 한국스피치커뮤니케이션학회(20건), 에세이코리아(25건), 한국고전의세계(51건), 한국현대문학회(97건)를 제외한 다른 열 개의 발행자로부터는 단지 한 건에서 열 건 안팎 정도의 문서밖에 수집되지 않았다. 하부주제 '문학' 부문의 23건 레코드 가운데는 1건만이 마광수가 발행자이지만, '한국문학' 부문의 588건 레코드 가운데 마광수가 발행자인 것은 모두 389건이다. 또한 발행자는 사회평론사, 해냄, 청하, 조선일보사 등으로 되어 있으나 실제 문서의 저작자는 마광수인 것까지 합치면 모두 401건이 마광수의 웹 문서이다.

저작자 및 발행자의 편향성 문제는 '사회과학' 분야에서도 나타나고 있다. '사회과학'은 모두 10,519건의 웹 문서를 포함하고 있는 주제 분야로서, 32,566건의 웹 문서를 수집하고 있는 '기술과학' 분야 다음으로 방대한 규모이다. '사회과학' 분야는 사회과학, 경제학, 정치학, 법학, 풍속・민속학, 통계학, 사회학・사회문제, 행정학, 교육학 및 국방, 군사학의 열 개 하부주제로 세분되어 있다. 〈표 5〉는 '사회과학' 분야에서 네 개 하부주제 분야 웹 문서 건수, 발행자, 수집일 등을 보여주고 있는데, 여기에서도 '문학' 분야와 마찬가지로 발행자의 편향성이 나타나고 있다. 특히 '사회과학' 분야에서는 소수의 발행자에 집중되어 있을 뿐만 아니라, 기관보다는 개인 발행자에게로 편향되어 있기도 하다.

예를 들어, 모두 203건의 웹 문서가 수집되어 있는 법학 분야에서는 한국해양대학교 김영구 교수(48건), 김종보 교수의 건축법교실(25건), 임지봉, 표창원 등과 같이 개인이 생산한 자료가 많은 비중을 차지하는 한편, 법 관련 주무 정부 부처인 법무부나 각급 법원, 입법기관인 국회 등에서 생산하였거나 그에 관련된 문서는 거의 보이지 않는다. 정부 부처인 문화체육관광부 자료는 '종교법인', '법인등록현황', '2003년

www.kci.go.kr

〈표 5〉 '사회과학' 분야 4개 주제 웹 문서의 발행자(저작자) 분석 [접근: 2009. 5. 27, 28; 6. 5]

주제	총문서	발행자(저작자)	문서	수집일
사회과학	16	충남발전연구원	2	2008. 4. 16
		대외경제정책연구원	2	2006. 4. 26
		한국외대아프리카연구소	4	2006. 3. 22
		유석춘	4	2005. 5. 31
		이론(김세균)	1	2005. 5. 27
		김상묵	2	2005. 5. 24
		한국첨단게임산업협회	1	2004. 10. 4
법학	203	한국건설감리협회	1	2007. 5. 10
		한국교육연구소	2	2007. 4. 2
		민족문제연구소	30	2007. 3. 28
		세계밀알협회	1	2006. 12. 12
		대외경제정책연구원	3	2006. 4. 18
		동서대학교일본연구센터	4	2006. 3. 30
		김종보교수의 건축법교실	25	2005. 12. 5
		이동복	1	2005. 11. 23
		국립공원관리공단	3	2005. 10. 28
		문화관광부	5	2005. 9. 20
		김영구	48	2006. 6. 16
		임지봉	11	2005. 6. 3
		표창원	14	2005. 6. 3
		명순구	2	2005. 5. 30
		한국노동연구원	1	2005. 5. 23
		정보통신정책연구원	31	2005. 4. 11
		한국여성개발원	7	2004. 10. 20
		대한민국학술원	1	2004. 10. 5
		한국첨단게임산업협회	2	2004. 10. 4
		청소년보호위원회	10	2004. 6. 18
		국방부	1	2004. 5. 17
풍속, 민속학	293	충남발전연구원	5	2008. 4. 30
		임재해	2	2006. 10. 30
		동서대학교일본연구센터	11	2006. 3. 30
		국립민속박물관	185	2006. 3. 23
		한국외대아프리카연구소	3	2006. 3. 22
		임재해	71	2005. 12. 6
		농지개량조합연합회	1	2005. 6. 21
		한국고전의세계	15	2004. 11. 10
통계학	6	충남발전연구원	1	2008. 4. 16
		국립공원관리공단	1	2005. 10. 28
		유석춘	1	2005. 5. 31
		한국어정보처리실	3	2004. 11. 19

www.kci.go.kr

문화관광부 법인현황', '2004년 법인현황' 등 5건의 웹 문서가 포함되어 있고, 국방부 자료가 1건이 포함되어 있을 뿐이다. 앞서 〈표 2〉에서 보는 OASIS의 '수집 우선 자료'에 "중앙정부가 생산한 온라인 디지털자원 및 대학간행물 등"이 열거되어 있음에도 불구하고, 실제로 수집된 웹 문서는 대부분 개인 저작자의 것들이다.

한편 기관 발행자 혹은 저작자의 권위 문제도 다소 고려해 볼 대상이라고 하겠다. '문학'과 '사회과학' 분야에서 가장 많은 웹 문서가 수집된 대학 기관은 동서대학교 일본문화센터와 한국외대아프리카 연구소이다. 동서대학교 일본문화센터는 한국문학(1건), 일본문학(1건), 법학(4건), 풍속・민속학(4건) 분야 웹 문서들의 발행자이며, 한국외대 아프리카연구소는 프랑스문학(1건), 기타 제문학(10건), 사회과학(4건), 행정학(1건) 분야의 발행자이다. 이것은 연구자의 주관적 견해일 수도 있으나, 우리나라에서 문학과 사회과학을 비롯한 다양한 학문 분야에서 많은 연구를 수행하고 있는 서울대학교, 한국학중앙연구원 등 주요한 교육, 연구 및 학술기관들이 발행한 웹 문서들은 전혀 보이지 않는 한편, 이들 두 개 대학 기관의 웹 문서만이 수집되어 있다는 점은 다소 의문을 야기한다. 이들이 적어도 웹 자원의 탁월성에 관한 한, 다른 기관들을 능가하는 대표적 기관이기 때문인가? 아니면 어떤 이유로든 이들이 단순히 웹 자원을 수집하기 쉬웠기 때문인가?

또한 개인 저작자의 권위에도 의문의 여지는 있다. 풍속・민속학 분야에서 수집된 "일본 가업의 계승 현황에 대한 연구 시니세를 중심으로(수집일 2006. 3. 24)"는 발행자가 세종연구소일본연구센터이지만, 실제 파일은 세종연구소 석・박사 세미나에서 발표된 전재은의 석사학위 논문이다.[3] 이 논문이 얼마나 중요한 "한국과 관련된" 학술적 혹은 문화적 가치를 지니고 있는지 평가할 수 있는 근거는 아무 데도 없으나, 제목 혹은 저작자의 신분 등으로 보아 과연 OASIS가 이 특정한 개인의 석사 논문을 국가적 디지털 자산으로 수집하고 보존해야 하는지 의문을 가질 수 있다고 하겠다.

4.3 정보의 유일성

OASIS가 '정보의 유일성'을 어떻게 정의하는 지는 어디에도 명시하지 않았다. 그러나 대체로 디지털 문화유산의 보존이 원래부터 디지털 형태로만 생산되었거나, 디지털이 아닌(non-digital) 동등한 버전이 없는 자원들을 대상으로 한다는 점을 고려할 때(UNESCO 2003, 30), OASIS 수집 콘텐츠에 '정보의 유일성'이란 기준이 적절히 적용되고 있는지에 대한 의구심을 갖게 한다.

먼저 '정보의 유일성'을 수집된 매체 형태 혹은 콘텐츠의 유일성 측면에서 볼 수 있다. 앞서 '문학' 분야에 수집된 웹 문서 가운데 마광수가 발행자이거나 저작자로서 관련된 웹 문서가 401건이라고 지적한 바 있다. 이 웹 문서들의 목록을 살펴보면 이전에 간행된 단행본인 '광마일기', '자궁 속으로', '즐거운 사라' 등의 제목을 볼 수 있다. 그 가운데 '광마일기'는 현재도 인터넷 서점에서 행림출판(1992)과 사회평론(1998)

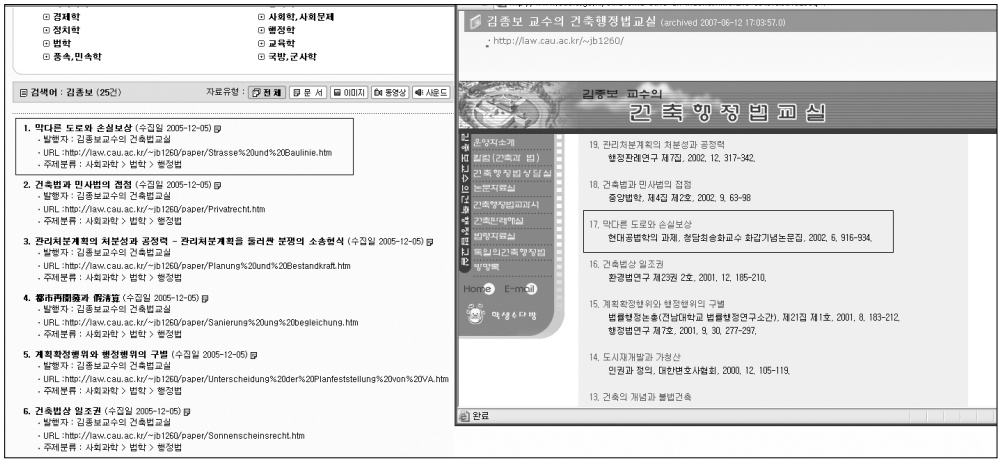
3) 〈http://sejongjapan.com/bbs/data/Policytalk/재은석사논문발표.doc〉.

www.kci.go.kr

간행본이 판매되고 있으며, '자궁 속으로'(사회평론 1998) 역시 판매되고 있다. '즐거운 사라'는 시중에서 구할 수 없기는 하지만, 여기에서는 단행본 소설의 각 장을 각각 하나의 웹 문서로 간주하여 수집해 놓은 것을 알 수 있다. 말하자면 이들은 어디서든 다른 매체, 즉 원래의 인쇄 매체 형태로 구할 수 있는 콘텐츠로서 정보의 유일성을 갖지 않는다.

'정보의 유일성'은 또한 수집된 웹 문서와 웹 사이트 간의 관계 측면에서도 볼 수 있다. OASIS는 분명 웹 문서와 웹 사이트를 구별된 자원으로 취급하며, 수집 자료 통계에서도 별도로 처리하고 있다. 그러나 현재 OASIS 콘텐츠를 살펴보면 이들은 중복되어 있다. 다시 말하면 OASIS는 수집된 웹 사이트에 포함된 각각의 웹 문서들을 별도로 수집한 자료인 것처럼 제시하고 있다. 물론 OASIS 수집 웹 문서와 웹 사이트를 전수(全數) 조사하지 않았으므로 모든 웹 자원이 그렇다고 단정할 수는 없으나, 적어도 이 연구에서 무작위로 살펴본 상당수의 웹 자원들이 그렇다.

예를 들면, '사회과학' 분야에 수집된 웹 사이트는 모두 73종으로, 그 가운데 '김종보 교수의 건축행정법교실'[4]이 포함되어 있다. 그리고 〈표 5〉에 보는 것처럼 '법학' 분야에 수집된 웹 문서에는 '김종보 교수의 건축행정법교실'이 발행자인 자료가 25건 수록되어 있다. 〈그림 2〉의 왼쪽 화면은 '김종보 교수의 건축행정법교실'이 발행자인 이들 웹 문서 25건의 리스트이고, 오른쪽 화면은 현재 아카이빙 되어 있는 그 웹 사이트를 보여준다. 왼쪽 화면의 리스트에서 첫 번째 웹 문서인 '막다른 도로와 손해 보상'은 2005년 12월 6일에 수집된 자료인데, 동일 문서가 2007년 6월 12일 아카이빙 된 이 웹 사이트에 수록되어 있다. 나머지 웹 문서들도 마찬가지이다. 다시 말하면, OASIS가 "하나 또는 몇 개의 파일 단위로서 정보의 가치를 가지게 되는 개별자원"으로서 2005년 수집하였다는 웹 문서들이 사실



〈그림 2〉 OASIS의 '김종보 교수의 건축행정법교실' [접근: 2009. 6. 5]

4) 〈http://law.cau.ac.kr/~jb1260/〉.

www.kci.go.kr

은 그보다 2년 뒤에 수집된 웹 사이트에 그대로 포함되어 있는 것들이다.

동일한 사례는 웹 문서 '법학' 분야에 2006년 6월 16일 48건이 수집된 발행자 '김영구'의 경우에서도 발견된다. URL로 확인해 본 이들 '김영구' 웹 문서의 출처는 '김영구 교수의 여해연구소'라는 웹 사이트[5]였고, 이것은 웹 사이트 '법학' 분야에 수집되어 2005년 8월 30일(아카이빙 버전1), 2007년 6월 12일(아카이빙 버전2)에 두 차례 아카이빙 되어 있다.

OASIS의 '자료 수집 절차'에 따르면 웹 문서와 웹 사이트에 대하여 별도의 선정 기준과 수집 절차 등이 있고, 반드시 "기존 수집 웹 자원과 중복되는지를 검사"하는 과정을 거친다고 하였다. 그러나 만약에 OASIS의 자료 수집 과정에서 동일 유형 자원 간에만 중복을 검사하고, 서로 다른 유형 즉, 웹 문서와 웹 사이트 간 교차 중복 검사를 수행하지 않는다면 이런 문제는 발생할 수도 있다. 또한 이전에 수집된 김종보 교수나 김영구 교수가 생산한 개별 웹 문서들이 매우 우수하였으므로, 이후에 아예 그들의 웹 사이트 자체를 수집했다고 해석할 수도 있다. 따라서 혹시 그동안 어떤 이유로든 수집 웹 자원의 통계치를 높이기 위하여 중복을 의도적으로 허용한 것이 아니라면, 전자의 경우는 향후 자원 간 교차 중복 검사를 시행하는 것만으로도 문제점을 해결할 수 있다고 보인다. 후자의 경우는 충분히 중복을 허용해도 무방하다고 하겠다. 그러나 그 이유와 해결책이 여하간 현재 이들은 '정보의 유일성'이라는 수집 기준에는 벗어나 있다.

4.4 학술적 내용

OASIS 수집 콘텐츠가 얼마나 학술적 내용을 갖는 것들인지 확인하는 데는 다소 어려움이 있다. OASIS '자원 수집 지침'에서 "국내의 저명한 대학 교수나 대학 연구원 등 권위 있는 한국 사람에 의해 쓰여지거나 국내외적으로 해당 학문분야에 기여를 한 것"이라는 조건을 갖고 있기는 하지만, 실제 수집된 자원들이 모두 그러한 조건을 만족하고 있는지를 직접 평가하기 어렵다는 의미이다.

물론 일부는 확인할 수 있다. 예를 들어 〈그림 2〉의 '김종보 교수의 건축행정법교실' 웹 사이트는 저작자가 교수라는 사실과 더불어 수록 문서들 각각에 대하여 학술지라는 출처를 밝히고 있다. 따라서 웹 사이트 자체의 학술적 가치를 어느 정도 짐작할 수 있고, 앞서 '정보의 유일성'과 관련하여 지적한 것처럼 이 웹 사이트에 수록되어 있지만 개별 웹 문서로 다시 중복 수집된 웹 문서들도 당연히 학술적 가치를 가진다고 할 수 있다.

그러나 웹 문서만으로는 그것을 알기 어려운 경우도 있다. OASIS에서는 웹 자원을 기술하기 위한 메타데이터 요소로 URL, 수집파일명, 저작자, 주제분류, 요약, 발행자, 생성일 및 UCI 등을 포함하고 있다. 그러나 이것들만으로는 종종 웹 문서의 발행자 혹은 저작자의 신원 혹은 권위를 알기 어렵고, 결과적으로 웹 문서 자체의 가치를 알기 어렵다는 문제점이 있다.

예를 들어 '문학' 분야에 수집된 웹 문서 가운데 "연세대학교 마광수 교수의 〈시 상징의 형이

5) 〈http://www.kocean.org/〉.

www.kci.go.kr

상성 연구〉에 관한 글"은 메타데이터 요소 가운데 요약을 통해 출처가 마광수교수의 개인 웹 사이트인 "http://www.makwangsoo.com."이라는 기본적 서지정보를 알 수 있으나, 실제 이 웹 문서가 어떤 학술적 가치를 갖는지 확인할 수는 없다. 메타데이터에서 이러한 정보를 얻을 수 없을 뿐 아니라, 실제로 열어본 이 웹 문서에서도 그것을 알 수 없었다. 이것은 한글(hwp) 파일로서 이것이 저자의 개인 웹 사이트 말고는 언제 어디에 발표되었는지, 어떤 형태로든 학술적 평가를 거친 것인지, 단순한 저자의 사견인지, 최초 버전인지 혹은 최종 버전인지 등을 알 수 없었다. 교수라는 저작자의 신분이나 제목을 보고 학술적 가치를 어느 정도 추측할 수는 있으나 그것만으로 충분하지는 않다.

다른 예를 들자면 '한국고전의세계'라는 발행자는 어른들 놀이, 경상도 손님굿 무가, 서울 손님굿 무가, 충청도 손님굿 무가 등을 채록한 문서 파일 등을 포함하여 한국문학 분야에서 51건의 웹 문서, 풍속, 민속학 분야에서 15건의 웹 문서 등을 제공하고 있는데, 이 발행자가 기관이나 단체인지, 책 표제인지 등 그 정체를 알 수 있는 정보가 어디에서도 주어지지 않는다. 다만 발행자가 '한국고전의세계'인 자료들의 URL 앞부분에 모두 '〈http://gojun.knu.ac.kr/...〉'가 붙어 있으므로, 이 도메인을 사용하는 경북대학교에 속한 기관 혹은 경북대학교와 관련된 발간자료일 것이라고 추측할 수는 있으나, OASIS 메타데이터만이 아니라 파일 자체에서도 확인할 수 없었다. 왜냐면 OASIS에 수집되어 있는 것은 개별 한글 파일로서 아무런 저작자나 발행자에 대한 정보를 수록하고 있지 않기 때문이다. 나중에 다른 경로로 확인한 결과 이 발행자, 즉 출처는 2003년 한국문예진흥원의 우수 웹 사이트로 선정된 적이 있는 "경북대 김문기 교수와 함께 하는 한국고전의 세계"라는 개인 웹 사이트임이 밝혀졌다. 더욱이 이 웹 사이트 자체는 문학 분야에서 수집된 11개의 웹 사이트들 가운데 하나였음에도 불구하고, 이처럼 별도로 수집된 웹 문서들에는 출처 정보가 전혀 언급되지 않았다.

4.5 정보의 최신성

'정보의 최신성'은 OASIS '자원 수집 지침'에서 명시하고 있는 기준 항목이다. 도서관이나 정보센터에서 어떤 자원의 컬렉션을 구성하든 '최신성'은 데이터의 품질, 정확성, 권위와 더불어 주요한 고려 대상이 되며, 특히 전자자원에서는 이러한 요소들이 더욱 중요하게 여겨진다(Evans 2000, 221). 일반적으로 '최신성'은 내용 및 입수의 최신성 면에서 판단할 수 있고, "단명"하고 "휘발성" 있는 웹 자원에 대해서도 마찬가지 기준이 적용될 수 있다.

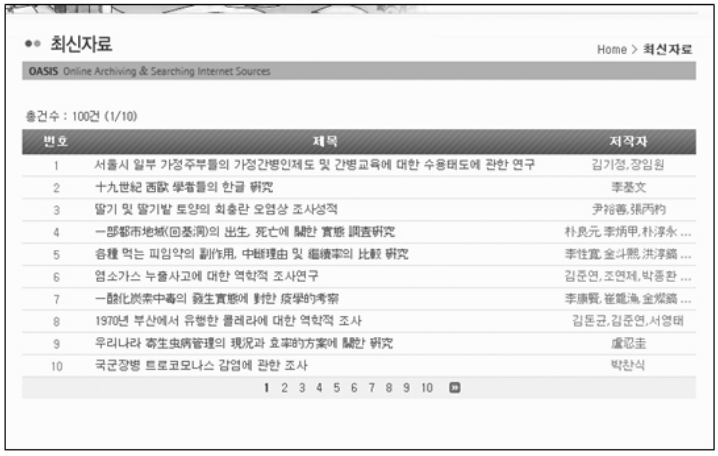
'정보의 최신성'은 앞에서 〈그림 1〉의 OASIS 웹 사이트의 초기 화면 왼쪽 하단에 제시된 '최신자료'를 중심으로 살펴보았다. '최신자료'는 웹 자원과 웹 사이트로 나누어서 제시되고 있는데, 여기에서 웹 자원은 OASIS의 다른 곳에서는 웹 문서라고 불리는 것이다. 이 연구를 위하여 OASIS 웹 사이트에 처음 접근할 당시(2009년 6월 5일) 초기 화면에서 '최신자료' 웹 자원은 "서울시 일부 가정주부들의 간병인제도 및 간병교육에 관한 수용태도에 관한 연구", "十九世紀 西歐 學者들의 한글 硏究", "딸기 및 딸기밭 토양의 회충란 오염상 조사성적" 및 "一部都市地

www.kci.go.kr

域(回基洞)의 出生, 死亡에 關한 實態 調査 硏究"가, 웹 사이트는 한국아메리카학회[6], 한국의류학회, 한국인구학회, 춘천국제연극제가 각각 열거되어 있었다.

먼저 웹 자원, 즉 웹 문서 '최신자료'의 상세한 내용을 검토하기 위하여 〈그림 3〉과 같은 '최신자료' 목록으로 들어간 결과, 각 웹 문서의 서지 레코드에서 제공되는 메타데이터 요소에는 URL, 수집파일, 저작자, 주제분류, 요약, 발행자, 생성일, UCI가 포함되어 있고, 수집 일자는 포함되지 않았다. '최신자료' 목록 맨 위에 열거된 "서울시 일부 가정주부들의 가정간병인제도 및…"(발행자: 대한예방의학학회)[7]의 생성일은 1988년 12월이다. URL 및 생성일 정보(http://data.healthis.org/pv/198802/a19.pdf)로 보아 이 문서는 지금부터 20년 전에 작성된 것으로서, 내용 상으로는 결코 '최신자료'일 수 없었다. 그처럼 내용 상 최신성을 갖지 않는다면 최근에 수집했다는 의미에서 '최신자료'라고 할 수는 있으나, 이를 뒷받침하는 수집일자의 정보는 제공되지 않는다. 이 자료를 검색화면에서 별도로 제목으로 검색한 결과의 메타데이터에서는 수집 일자가 2006년 3월 14일임을 확인할 수 있었는데, 이 연구가 진행되는 2009년 6월의 시점에서 이미 3년 전에 수집된 웹 문서를 '최신자료'라고 보기엔 무리가 있다.

그 다음에 열거된 "十九世紀 西歐 學者들의 한글 硏究"(발행자: 대한민국 학술원)는 생성일은 2000년 1월 1일, 이 자료의 제목 검색 결과 수집일자는 2004년 10월 5일이었다. "딸기 및 딸기밭 토양의 회충란 오염상 조사성적"(발행자: 대한예방의학학회, 생성일: 1971년 10월)과 "一部都市地域(回基洞)의 出生, 死亡에 關한 實態 調査硏究"(발행자: 대한예방의학학회, 생성일: 1971년 10월)도 마찬가지로, 결국 여기 제시된 자료들은 생성일로도 수집 일자로도 '최신자



〈그림 3〉 '최신자료' +More 리스트 [접근: 2009. 6. 6 오후 12:45]

6) 〈http://www.asak.or.kr/〉.
7) 〈http://data.healthis.org/pv/198802/a19.pdf〉.

www.kci.go.kr

료'라고 하기엔 어려운 자료들임이 판명되었다.

'최신자료'의 변동 여부를 검토하기 위하여 최근 다시 검토한 웹 문서 '최신자료' 목록 상위 10건 가운데 레코드 번호 1부터 5까지 5건의 웹 문서는 2009년 6월 이후 추가된 새로운 자료이지만, 계속되는 검색 오류로 수집 일자를 확인하기 어려웠다(접근: 2009. 10. 15 - 17). 이 웹 문서들의 생성일은 각각 1: 미상, 2: 2005, 3: 2007. 12, 4: 2006. 12, 5: 2004, 6: 1988. 12, 7: 2000. 1, 8: 1971. 10, 9: 1971. 10 그리고 10: 1970. 10이다. 또한 '최신자료' 목록에서 살펴본 웹 문서 11부터 13은 1971년, 14부터 24는 1979년, 25는 1980년, 25부터 30은 1978년에 생성된 자료들이다. 웹 문서 8부터 30까지는 모두 대한예방의학학회가 생성한 웹 문서들로 2006년에 입수되었다. 수집 일자 확인이 어려운 레코드 번호 1부터 5까지의 웹 문서를 제외한 나머지는 모두 2006년 무렵에 입수된 것들이다. 그 가운데 1980년에 생성된 웹 문서 25를 제외한 나머지 대부분이 적어도 30여 년 전에 생성되었고 나중에 pdf 파일로 변환된 텍스트 자료들임을 알 수 있다.

5. 맺음말

OASIS는 국립중앙도서관이 구축하여 운영하고 있는 국가 디지털 자원 아카이브이다. OASIS 구축과 운영 초기 당시 여러 연구자들이 우리나라 국가대표도서관으로서 국립중앙도서관의 디지털 유산 수집과 보존을 위한 정책 수립 및 실행 방안에 대하여 많은 것을 제안하였다. 특히 최원태(2005)는 디지털 유산 보존의 국가적인 해결책은 "포맷에 관계없이 역사적으로 중요한 문화 자원과 정보자원을 계속적으로 수집, 조직, 관리하는 것"임을 강조한 바 있고, 그보다 앞서 이혜원(2004, 8)은 국립중앙도서관이 이미 "우리나라를 대표하는 행정, 교육, 정치, 문화, 사회 기관에서 생산된 온라인 디지털자원을 확보하기 위한 기준 및 방안을 제시하고 실행하는 과정에 있으며… 현 시대에서 중요하게 다루어지는 이슈들과 인지도 높은 기관과 개인이 제작한 자료 및 국제적인 회의, 전자저널까지 현재, 미래에 유용한 정보로 쓰일 자원들을 수집하여 관리하고 있다"고 주장하였다. 그러나 불과 사오년이 지난 지금 OASIS의 현황은 전자가 "그래야 한다"고 제안한 방향으로도 가고 있지 않으며, 후자가 "이미 그렇게 하고 있다"고 주장한 것처럼 실행되고 있지도 않는 것으로 보인다.

2009년 5월 20일부터 7월 2일 사이, 그리고 다시 10월 5일부터 17일 사이 이 연구에서 살펴 본 OASIS는 그동안 수집된 콘텐츠에 관련된 몇 가지 문제점이 있는 것으로 나타났다. 이들을 OASIS '자원 수집 지침'에서 제시된 정보자체의 특성에 따라 정리해 보면, 주제 분포의 편중, 저작자/발행자 편향성과 권위의 근거 미약, 정보의 유일성과 최신성 결여, 웹 문서와 웹사이트의 중복 수집, 학술적 가치의 근거 결여 등을 들 수 있다. 말하자면 OASIS 콘텐츠가 '자원 수집 지침'에 명시된 기준에 부합하지 않음을 의미한다. 이 연구에서 '문학'과 '사회과학' 분야를 중심으로 살펴보았기 때문에 이러한 문제점들이 단지 두 분야에만 국한되는 것일 수도 있다. 그렇다고 해도 이렇게 지적된 문제점에 대하여 국립중앙도서관이 스스로 보다 심층

www.kci.go.kr

적 분석과 더불어 구체적 개선방안을 마련해야 할 것으로 보이지만, 여기에서 두 가지 개선방안을 제안하도록 한다.

첫째, OASIS 콘텐츠 수집에 개입하는 사람들의 역할과 수준을 강화해야 하고, 그러기 위해 웹 자원 추천자 "실명제"를 시행할 필요가 있다.

현행 OASIS의 자원 수집 업무는 기본적으로 5단계로 구분되어 있고, 그 가운데 첫 단계인 '수집대상 디지털자원의 선정단계'에서 국립중앙도서관 수집팀, 각 주제 분야별 전문가 및 일반인이 수집 대상 자원의 추천에 개입한다고 하였다(국립중앙도서관 2009, 145). 이를 위해 인문과학, 사회과학, 자연과학 및 정책자료분과에 분야별 전문가 15명 내외로 구성된 '온라인디지털자원수집보존위원회'가 있고, 이들이 "인터넷의 가치 있는 온라인 디지털자원의 체계적 수집 · 보존을 위한 중요사항의 심의 및 자문"을 수행한다고 하였다(국립중앙도서관 OASIS 2009b). 또한 초기에는 "학술적으로 가치 있는 자원 수집을 위해… 인문, 사회, 자연과학 28개 주제 분야의 7,000여개 URL을 추천받았으며, 각 주제의 주제전문가 84명을 대상으로 인력풀을 구성" 하였음을 강조하였다(이혜원 2004, 15). 이처럼 적지 않은 수의 주제전문가가 실제로 OASIS의 웹 자원 수집 과정에 지금도 개입하고 있다면, 누가 어떤 웹 문서나 웹 사이트를 어떤 이유로 추천하였는지, 그 웹 자원이 과연 수집 조건에 부합하는지, 실명으로 타당하고도 투명한 추천의 근거를 제시할 수 있는 정도의 책임을 가져야 할 것이다.

둘째, OASIS에 수집된 웹 자원의 수집 근거를 명시해야 하고, 그러기 위해 메타데이터 요소를 확장할 필요가 있다.

현재 OASIS는 웹 자원을 기술하기 위해 URL, 수집파일명, 저작자, 주제분류, 요약, 발행자, 생성일 및 UCI 등의 메타데이터 요소를 사용하고 있다. 그러나 그것만으로는 수집된 웹 문서 혹은 웹 사이트가 어떤 가치를 가지며, 어떤 근거로 수집되어 있는지 알기 어렵다. 다른 나라의 예를 들자면 일본 국립국회도서관의 웹 아카이브인 WARP(日本. 國立國會図書館 2009b)은 메타데이터 ID, 서지표제, 기타표제, 편자(編者), 출판자 · 공개자, 공개일, 기점(起点) URL, 과거의 기점 URL, ISSN, NDC(日本十進分類法) 분류, 컬렉션종별(種別), NDL자원유형, 주석 등의 서지정보와 수집일 범위, 본문, 페이지제목, URL, 도메인, 파일 종별 등의 본문정보 등 매우 상세한 정보를 제공해 주고 있다. 물론 WARP도 수집된 웹 자원이 "국가적 자산"으로서 어떤 가치를 가졌는가를 직접 명시하지는 않는다. 그러나 적어도 '컬렉션종별산'기점에서 해당 웹 자원이 국립국회도서관이 정한 WARP의 컬렉션 대상, 즉 국가기관, 도도부현, 정령지정도시, 시정촌합병, 법인 · 기구, 대학, 이벤트, 전자잡지, 기타 등의 범주에서 어디에 속하는지를 명시해 줌으로써 선정기준에 합당한 자료임을 보여준다.

OASIS에서도 그와 마찬가지로 특정한 웹 자원이 국립중앙도서관이 제시하고 있는 수집 조건의 무엇에 해당하는지, 혹은 '수집우선자료'의 어떤 범주에 속하는지 메타데이터의 한 요소로 명시해 줄 수 있다. 이를테면 '학술적 내용' 때문에 선정되었는지 '최근 이슈가 되는 온라인 디지털 자원'인지, 중요한 국제 '회의자료'인지 등을 명시하는 것이다. 만약 새로운 메타

www.kci.go.kr

데이터 요소를 추가하기 어려우면, 현재의 '요약' 요소에 이러한 내용을 추가 기재해 줄 수도 있을 것이다. 그럼으로써 디지털 문화유산으로서의 수집과 보존 근거를 보여주는 것이 바람직하다. 여기에서 먼저 제안한 추천 실명제가 추천자의 책임감을 강화하는 것이라면, 메타데이터 요소 확장은 이용자가 추천된 웹 자원의 가치를 인정하고 동의할 수 있게 하는 것이라고 하겠다. 물론 이러한 개선방안이 앞서 지적한 여러 문제들에 대한 직접적이며 즉각적인 해결책은 아니다. 그러나 이미 설정해 놓은 자원 수집의 기준과 지침이 있는 만큼, 그 기준에 철저히 부합하는 웹 자원을 전문가 책임으로 선정하고, 이용자들도 납득하고 신뢰할 수 있는 수집 정보를 제시하라는 것이다. 처음부터 정해진 기준대로 웹 자원을 선정하고 검수한다면 품질문제가 생기지 않을 수 있다.

삼십 여 년 전 미국 University of Pittsburgh 연구팀은 도서관 자료의 이용을 연구하면서 도서관들이 "쓸모 있으면서 이용되기도 할 '적절한(right)' 자료를 고른다는 것은 마치 건초더미에서 바늘을 찾는 것과 같다"고 하였다. 그들이 언급한 건초더미는 당시까지 구텐베르그 이래 발행된 3천만 종의 책과 세계적으로 간행되는 오만 내지 십만 종의 정기간행물, 1978년 한 해에 세계에서 간행된 오십만 권 가량의 책 등을 의미하였다(Kent 1979). 그와 비교할 수 없을 만큼 엄청난 속도와 분량으로 생산 및 유통되며 순식간에 사라지기도 하는 웹 자원 가운데 문자 그대로 "적절한" 자료를 수집하여 국가 디지털 자원으로 보존한다는 것은 "건초더미에서 바늘을 찾는 것"보다 더한 노력과 책임감이 필요한 작업이라고 할 수 있다. 또한 반드시 "바늘"을 찾을 수 있다는 보장도 없다. OASIS가 해야 하는 일이 아마도 그럴 것이다. 그래서 어렵기도 하고, 많은 격려가 필요하기도 할 것이다. 그러나 다른 나라도 하니까 우리도 하는 사업, 혹은 단기간에 수치 상 목표 달성만을 지향하는 사업이 아니라, 국가 디지털 문화유산을 수집하고 보존하여 다음 세대에까지 물려주겠다는 중차대한 목표를 가진 사업임을 기관 스스로 깊이 인식하며, 지향할 바를 수시로 점검하고 확인해야 할 것이다.

참 고 문 헌

[1] 국립중앙도서관. 2009. 『2008 국립중앙도서관연보』. 서울: 국립중앙도서관.

[2] 국립중앙도서관. 2009a. "OASIS 소개. 개요." [online]. [cited 2009.5.22]. 〈http://www.oasis.go.kr/intro/intro_overview.jsp〉.

[3] 국립중앙도서관. 2009b. "OASIS 소개. 수집보존프로세스." [online]. [cited 2009.5.30]. 〈http://www.oasis.go.kr/intro/intro_gatherprss.jsp〉.

[4] 국립중앙도서관. 2009c. "OASIS 소개. 추진경과." [online]. [cited 2009.5.22].

www.kci.go.kr

〈http://www.oasis.go.kr/intro/intro_history.jsp〉.
[5] 김유승. 2007. 웹 아카이빙의 법·제도적 문제에 대한 고찰: 웹 정보자원의 특성을 중심으로. 『한국문헌정보학회지』, 41(3): 5-24.
[6] 김유승. 2008. 복합적 웹 아카이빙 정책에 관한 고찰: 프랑스국립도서관의 사례를 중심으로. 『한국문헌정보학회지』, 42(4): 159-179.
[7] 서혜란. 2004. 『디지털 납본제도 방안』. 서울: 국립중앙도서관.
[8] 이소연. 2008. 믿을 수 있는 디지털 아카이브 인증기준: OASIS 적용사례. 『정보관리학회지』, 25(3): 5-25.
[9] 이혜원. 2004. 『온라인 디지털 자원 구축 사례: 국립중앙도서관을 중심으로』. 서울: 국립중앙도서관.
[10] 최원태. 2005. 『디지털유산 보존과 국가대표도서관의 정책방안』. 서울: 국립중앙도서관.
[11] Day, Michael. 2003. "Preserving the Fabric of Our Lives: A Survey of Web Preservation Initiatives." [online]. [cited 2009.6.12].
〈http://www.ukoln.ac.uk/metadata/presentations/ecdl2003-day/day-paper.pdf〉.
[12] Evans, G. Edward, & Saponaro, Margaret Z. 2000. *Developing Library and Information Center Collections*. 4th Edition. Greenwood Village, Colorado: Libraries Unlimited.
[13] Grotke, Abbie. 2008. *International Internet Preservation Consortium: 2008 Member Profile Survey Results.*. Canberra: The National Library of Australia.
[14] Kent, Allen, et al. 1979. *Use of Library Materials: The University of Pittsburgh Study*. New York: Marcel Dekker, Inc.
[15] Library of Congress. 2009. "Web Capture." [online]. [cited 2009.6.8].
〈http://www.loc.gov/webcapture/〉.
[16] Library of Congress Web Archives. 2009. "Minerva." [online]. [cited 2009.5.30].
〈http://lcweb2.loc.gov/diglib/lcwa/html〉/lcwa-home.html〉.
[17] Lyman, P. 2002. "Archiving the World Wide Web." *In: Building a National Strategy for Digital Preservation. Council on Library and Information Resources and Library of Congress.*, 38-51. [online]. [cited 2009.6.12].
〈http://www.clir.org/pubs/reports/pub106/web.html〉.
[18] National Library of Australia and Partners. 2009a. "PANDORA. Australia's Web Archive. PANDORA Overview." [online]. [cited 2009.5.30].
〈http://pandora.nla.gov.au/overview.html〉.
[19] National Library of Australia and Partners. 2009b. "PANDORA. Australia's Web Archive. PANDORA Archive Size and Monthly Growth." [online]. [cited 2009.5.30].
〈http://pandora.nla.gov.au/statistics.html〉.

www.kci.go.kr

[20] National Library of Australia and Partners. 2009c. "PANDORA. Australia's Web Archive. Selection Guidelines." [online]. [cited 2009.5.30]. 〈http://pandora.nla.gov.au/guidelines.html〉.

[21] UNESCO. 2003. "Guidelines for the Preservation of Digital Heritage." [online]. [cited 2009.6.12]. 〈http://unesdoc.unesco.org/images/0013/001300/130071e.pdf〉.

[22] 武田和也. 2008. 海外動向との對比からみた日本のWebアーカイビングの課題と展望-國立國會図書館の取り組みを通して. 『情報の科學と技術』, 58(8): 394-400.

[23] 國立國會図書館. 2009a. "本事業のご紹介." [online]. [cited 2009.5.22]. 〈http://warp.ndl.go.jp/WARP_Intro.html〉.

[24] 國立國會図書館. 2009b. "利用方法." [online]. [cited 2009.5.21]. 〈http://warp.ndl.go.jp/WARP_help.html#idx1-2-2〉.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] The National Library of Korea. 2009. *Annual Report 2008 National Library of Korea.* Seoul: The National Library of Korea.

[2] The National Library of Korea. 2009a. "About OASIS." [online]. [cited 2009.5.22]. 〈http://www.oasis.go.kr/intro/intro_overview.jsp〉.

[3] The National Library of Korea. 2009b. "About OASIS. Selection Target & Processes." [online]. cited 2009.5.30]. 〈http://www.oasis.go.kr/intro/intro_gatherprss.jsp〉.

[4] The National Library of Korea. 2009c. "About OASIS. History." [online]. [cited 2009.5.22]. 〈http://www.oasis.go.kr/intro/intro_history.jsp〉.

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

[6] You-Seung Kim. 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.

[7] Hye-Ran Suh. 2004. "The Operation of Digital Deposits." Seoul: The National Library of Korea.

[8] So-Yeon Lee. 2008. "Trustworthy Repositories Audit Criteria: Self-Assessment of OASIS." *Journal of the Korean Society for Information Management*, 25(3): 5-25.

[9] Hye-Won Lee. 2004. "Construction of Online Digital Resources." Seoul: The National

www.kci.go.kr

Library of Korea.

[10] Won-tae Choi. 2005. *Preservation of Digital Heritage and the Strategic Directions of the National Library of Korea*. Seoul: The National Library of Korea.

[22] Takeda, Kazuya. 2008. "Issues and Prospects about Web Archiving in Japan Compared with the Ttrends of Overseas: the Project of National Diet Library(NDL)." *The Journal of Information Science and Technology Association*, 58(8): 394-400.

[23] National Diet Library, Japan. 2009a. "Introduction to WARP." [online]. [cited 2009.5.22]. 〈http://warp.ndl.go.jp/WARP_Intro.html〉.

[24] National Diet Library. Japan. 2009b. "How to Use." [online]. [cited 2009.5.21]. 〈http://warp.ndl.go.jp/WARP_help.html#idx1-2-2〉.

www.kci.go.kr

www.kci.go.kr