과학기술정보유통을 위한 OAI기반 지식정보 수집·저장 시스템 구축

윤준원°, 이용식, 이상기, 신기정
한국과학기술정보연구원
{jwyoon°, yslee, sklee, kjshin}@kisti.re.kr

Construction of the Harvest and Repository system based on OAI for Circulation of Science Technology Information

Junweon Yoon°, Youngsik Lee, Sangki Lee, Kijung Shin
Korea Institute of Science and Technology Information

요 약

디지털 정보의 비중과 그에 대한 의존도가 현격이 높여져감에 따라 디지털 정보자원의 효율적인 수집, 저장 방법의 논의가 활발해져가고 있다. 이런 디지털 라이브러리의 분야에서 학술정보의 자유로운 이용(Open Access)에 대한 개념은 확고한 기반을 다져가고 있으며, 기술적 지원으로 제공된 OAI 프로토콜을 활용하여 해외의 학술자료들을 수집, 저장하여 배포하는 과학기술정보 유통체제는 이미 큰 이슈로 자리 잡고 있다. 한국과학기술정보연구원(KISTI)은 과학기술중앙정보센터로서 논문, 특허, 연구보고서, 사실정보, 생물다양성정보 등을 비롯한 다양한 종류의 과학기술관련 데이터베이스를 구축, 수집하여 서비스 하고 있다. 이에 국내외 과학기술 정보의 종합포털 및 디지털 전자유통 체제를 확립하기 위해 OAI 프로토콜을 통한 학술자료의 유통이 요구된다.

본 논문에서는 OAI 기반 지식정보 수집·저장시스템인 ***stOAI***(OAI 과학기술유통시스템)을 구축, 개발하였다. 본 시스템은 해외학술자료들을 OAI 프로토콜을 통해 수집·저장하여 Yeskisti(과학기술정보포털서비스)의 OA(Open Access)를 통해 해외저널을 무료로 배포하며 또한, KISITI의 과학기술정보를 효율적으로 외부 정보서비스 연계기관에게 제공하게 된다.

1. 서 론

OAI는 디지털 콘텐트를 보다 효율적으로 유통하기위해 유용한 각종 상호운용성 표준(interoperability standards)을 개발하고 보급하는 도구 역할을 수행한다[1]. 또한 OAI(Open Archives Initiative)는 메타데이터 수확(harvesting)을 위한 프로토콜로서 공식적으로 OAI-PMH(Protocol for Metadata Harvesting)라는 명칭을 사용한다.

OAI-PMH(이하 OAI 프로토콜)는 확장성이 뛰어나고 간결하며 국제표준인 HTTP, XML, Dublin Core 기술에 기반을 둔다.

OAI라는 용어가 의미하는 내용을 살펴보면 OAI의 Open은 제한 없는 사용 즉, 다양한 아카이브의 콘텐트 접근(공유)이 가능함을 의미하며, Archive는 디지털콘텐트의 저장소를 의미하며 레파지토리(Repository)로 사용되기도 한다. 이는 디지털 학술자원을 생산하는 대학이나 연구소 그리고 지적자산을 수집, 보전, 서비스를 하는 기관저장소를 포함한다. Initiative는 디지털 라이브러리 연합(Digital Library Federation)과 멜론 재단(Mellon Foundation)의 협력 프로젝트를 의미한다.

본 논문에서는 OAI 기반으로 하는 지식정보 수집 · 저장시스템인 ***stOAI***(OAI 과학기술유통시스템)을 구축, 개발하였다. 본 시스템은 OAI 프로토콜을 통해 약 60만 건의 해외 학술저널을 수집·저장하여 Yeskisti(과학기술정보포털서비스)의 OA(Open Access)를 통해 해외학술저널을 무료로 제공하고 있다.

2. OAI 개념

▪ 모델

OAI 프로토콜 모델은 크게 서비스제공자인 SP(Service Provider)와 데이터제공자인 DP(Data Provider) 두 개의 기능으로 구분된다[2][3]. DP는 디지털 정보를 수집, 보유하고 있는 시스템이며, SP에게 적합한 메타데이터를 표준적인 방법으로 제공한다. SP는 DP들로부터 수집한 메타데이터를 기반으로 부가가치(검색, 원문제공 등)정보를 서비스하는 역할을 수행한다.

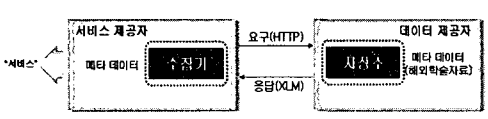

그림 1. OAI 기본 개념

▪ 프로토콜

OAI 프로토콜은 6개의 요청(명령) - GetRecord, Identify, ListIdentifiers, ListMetadataFormats, ListRecords, ListSets - 으로 구성되며 HTTP GET 또는 POST방식을 사용하여 메타데이터를 요청하게 된다. 모든 OAI 프로토콜 요청에 대한 응답은 XML 인스턴스(문서)로 인코딩된다[2].

	요청(명령)	기능 설명
저장소(아카이브)에 관한 정보를 얻기 위한 명령	Identify	저장소의 기본정보 (BaseURL등)를 검색
	ListSets	저장소의 셋(set)구조를 검색
	ListMetadata Formats	저장소에서 제공되고 있는 메타데이터 포맷 검색
메타 데이터 수집을 위한 명령	ListIdentifiers	저장소에서 수집 가능한 식별자(identifier) 검색
	GetRecord	저장소로부터 하나의 레코드를 수집
	ListRecords	한꺼번에 대량의 레코드를 수집

표 1. OAI 프로토콜 6가지 요청과 기능

▪ 메타데이터

각각의 저장소(Archive)는 자신의 응용영역에 적합한 요소들로 구성된 메타데이터를 사용할 수 있다. 그러나 OAI 프로토콜에 따라 메타데이터를 전송하기 위한 최소한의 부분집합 즉, 공통 포맷을 지원하여야 한다. OAI프로토콜 v1.0에서는 공통의 메타데이터 표준으로 기본 DC(Dublin Core)를 선택하였으나 반드시 기본 DC 포맷이 아니더라도 응용 영역에 적합한 메타데이터 표준을 사용하면 된다[2].

3. 적용 사례

영국, 네덜란드, 독일 등과 같은 정보선진국들은 학술정보를 표준적인 방식으로 유통하기 위한 OAI기반 학술지식정보 유통체계를 구축하고 있으며, 현재(2006년 4월) 400여개의 저장소가 OAI공식 홈페이지(http://www.openarchives.org)에 등록, 확대되어지고 있다. 국내에서 또한 KERIS(한국교육학술정보원)에서 OAI를 통해 국내 대학 및 교육 유관기관이 생성하는 학술자원을 수집, 통합, 서비스하는 시범사업을 진행하고 있다.

▪ ***AmericanSouth* (http://americansouth.org)**
AmericanSouth.Org는 Emory 대학에서 진행중인 프로젝트로 연구자원에 대한 메타데이터의 색인과 특정 분야(남미지역의 인문, 사회, 자연 과학, 역사학)에 대한 가치 있는 자원을 도서관과 박물관으로부터 수집하고 있으며, 통합, 색인, 검색을 위해 참여기관으로부터 주기적으로 메타데이터를 수확하여 중앙집중형 메타데이터를 구축, 서비스하고 있다[4][5]. 현재 31개의 기관으로부터 수집한 55,488개의 레코드들을 가지고 있다.

▪ ***OAIster* (http://www.oaister.org)**
*Andrew W. Mellon*재단의 후원으로 시작된 Michigan 대학의 디지털 라이브러리 구축 프로젝트로 디지털자원의 자유로운 이용을 목적으로 시작되었다. 초기 두 해는 *Illinois* 대학에서 제공한 수집기를 사용하였으며, 이후 *UIUC*와 공동으로 개발한 수집기를 사용하여 규칙적으로 수확된 데이터들을 가공, 제공하고 있다. 이렇게 수집한 데이터들은 사용자들에 학술정보의 자유로운 접근을 용이하게 한다[4]. 현재(2006년 3월)601개 기관으로부터 수집한 7,031,783건의 메타 데이터 레코드의 검색 서비스를 제공하고 있다.

▪***dColletion* (http://www.dcollection.net)**
*dCollection*은 국가학술연구DB 구축사업의 일환으로 KERIS(한국교육학술정보원)에서 시작된 "분산된 학술연구정보의 효율적인 통합관리 및 공동 활용을 위한 생성 및 유통체계 시스템"이다. 교육학술 분야에선 처음으로 추진된 시범 사업으로 개발한 디지털 아카이브 시스템은 학술정보 생산기관인 대학에서 자료수집과 동시에 수집한 자료를 즉시 유통시킬 수 있도록 하고 있다. KERIS는 학술자료를 생산하는 참여기관이 직접적으로 메타데이터와 원문을 구축하는 아카이빙센터 기능을 수행할 수 있도록 '학술자료구축 시스템'을 보급하고, KERIS는 중앙에서 메타데이터와 원문의 URL정보만을 수집하여 통합서비스를 제공하도록 설계하였다. 2005년 20개의 대학이 참여하고 있다[6].
한국과학기술정보연구원은 과학기술중앙정보센터로 디지털콘텐트(학술자료)의 효율적인 유통환경과 상호운용성을 증진시키기 위해 OAI 프로토콜을 이용한 디지털 수집 · 저장 시스템이 요구된다.

4. *stOAI* 프레임웍

*stOAI*는 OAI 프로토콜을 이용한 과학기술유통시스템으로 수집(Harvester)시스템과 저장(Repository)시스템으로 구성된다. 본 시스템에서 OAI 프로토콜을 이용하여 수집, 저장된 콘텐트(해외저널)는 과학기술정보 포털 서비스(www.yeskisti.net) 논문검색의 OA(Open Access)메뉴에서 무료로 제공하게 된다.

▪ 수집시스템(Harvester)

그림 2는 수집 시스템 모델을 보여주고 있다.

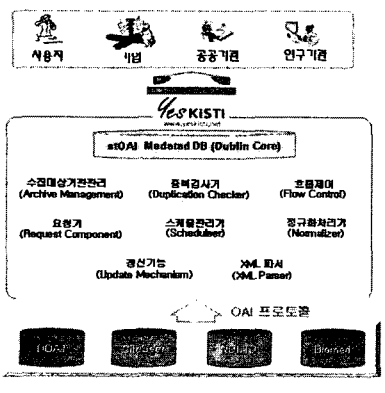

그림 2. *stOAI* 수집시스템 모델

수집시스템은 그림2와 같이 외부 데이터저장소로부터 메타데이터를 수집하여 *stOAI* 수집시스템에 구현된 기능모듈에 따라 XML로 인코딩된 데이터를 처리하게 된다. 표 2는 *stOAI* 수집시스템의 기능을 나타내고 있다.

기능	설명
수집대상기관관리	수집대상 DP를 관리
중복검사기	서로 다른 DP에서 수집한 데이터 중 중복 레코드를 처리
흐름제어	대량의 레코드를 일정한 크기로 분할하여 흐름제어 토큰(resumption token) 속성을 사용하여 전송
요청기	명시된 HTTP 요청을 생성하여, OAI DP로 전송하는 기능
스케줄관리기	DP를 작동시키기 위한 스케줄(범위,유형,갱신주기,처리상태,운영상태 등) 관리
정규화처리기	서로 다른 메타데이터 포맷으로 표시된 데이터를 동일한 구조로 변환하는 기능
갱신기능	기존에 수집한 메타데이터를 새로 수집한 메타데이터로 갱신하는 기능
XML 파서	수집한 XML로 인코딩된 데이터를 내부 데이터 구조로 변환하는 기능

표 2. *stOAI* 수집시스템의 기능

그림3은 *stOAI* 수집시스템의 프로세스를 보여주고 있다.

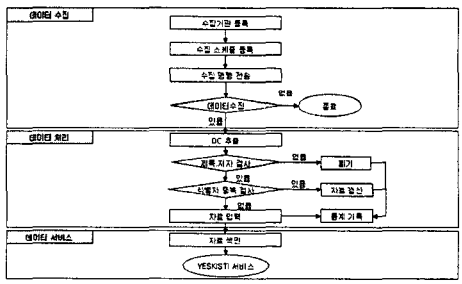

그림 3. 수집 프로세스

데이터 수집과정에서는 먼저 BaseURL를 통해 수집할 기관의 데이터저장소(2006년 3월 현재 www.openarchives.org 내에 400여개의 OAI 저장소가 등록되어있다.) 정보를 가져온 후, 이를 등록하게 된다. 그림 4은 OAI 프로토콜의 Identify의 속성을 통해 해당 XML로 인코딩된 데이터저장소의 기본정보를 나타내고 있다.

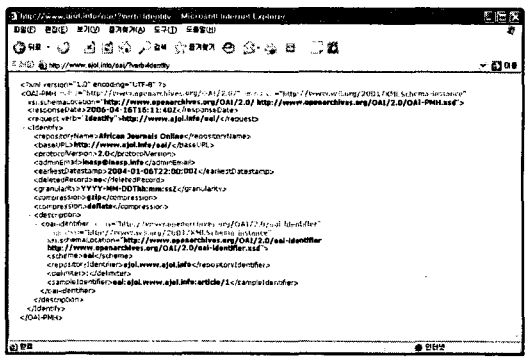

그림 4. 데이터저장소 응답레코드

수집기관의 정보가 등록이 되면 등록 기관의 수집스케줄 정보(수집시작·종료일, 수집주기, 처리상태, 운영상태, set관리)를 등록한다. 수집스케줄에 따라 수집명령을 전송하고 해당되는 저장소로부터 데이터를 수집하게 된다. 데이터 처리과정에서는 DKRS는 수집된 데이터를 DC 포맷으로 저장하게 되므로 각 저장소로부터 수집된 메타

데이터를 DC포맷에 맞게 추출하게 된다. 이렇게 추출된 데이터 중 제목, 저자와 같은 기본 항목이 누락된 데이터들은 폐기하고, 식별자가 중복된 경우에는 기존 정보를 갱신하게 된다. 마지막으로 정제된 데이터를 입력하고, 데이터 서비스과정에서는 저장된 데이터들을 색인과정을 통해 Yeskisti(과학기술정보포털서비스)에서 서비스하게 된다.

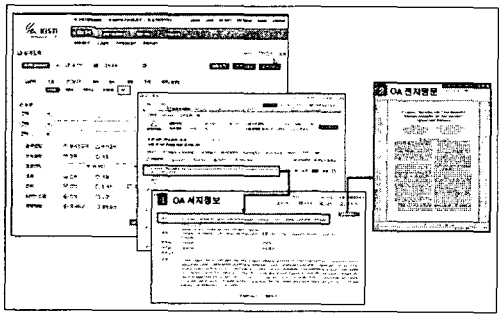

그림 5 Yeskisti의 OA(무료저널) 서비스

▪ 저장시스템(**Repository**)

stOAI 저장(Repository) 시스템에서는 학술자료(학위논문, 학술지논문, 연구보고서 등)를 자체적으로 생성(제출)하거나, KISTI에서 보유하고 있는 논문, 특허, 연구보고서, 사실정보, 생물다양성정보 등을 비롯한 다양한 종류의 과학기술관련 데이터베이스를 OAI 프로토콜을 이용해 외부 정보서비스 연계기관(네이버, 다음, 엠파스, 한국정보문화진흥원 등)에 제공하게 된다. 그림 6은 *stOAI* 저장시스템 모델을 보여주고 있다.

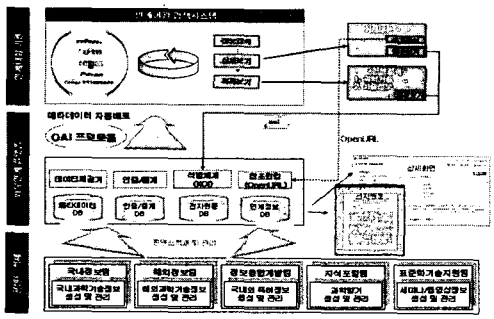

그림 6. *stOAI* 저장시스템 모델

그림 7은 저장 프로세스를 나타내고 있는데 과학기술정보 메타데이터를 OAI 프로토콜을 이용 XML로 인코딩하여 제공하게 된다.

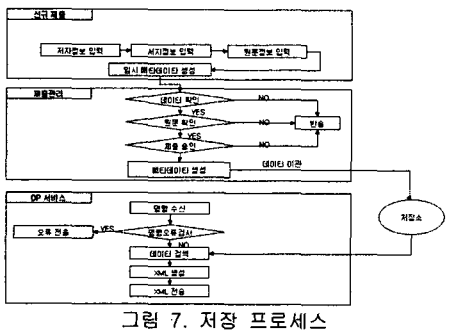

그림 7. 저장 프로세스

신규제출은 학술정보 생산자로부터 제출정보(저자, 서지정보, 원문 등)를 입력받아 임시 메타데이터를 생성하게 된다. 제출관리에서는 임시 메타데이터 정보를 정제하는 과정으로 저장된 데이터와 원문을 검사한 후 승인과정을 통해 실재 메타데이터를 생성하고 원문 서비스를 위해 저장소에 데이터를 이관하게 된다.

DP서비스에서는 수집시스템의 명령(요청)을 받아 저장소에 있는 메타데이터를 검색하여 적절한 정보를 XML 인스턴스로 인코딩하여 제공하게 된다. 그림 8은 요청에 따라 OAI 프로토콜 중 GetRecord를 통해 stOAI 시스템으로부터 하나의 레코드를 XML 데이터로 보여주고 있다.

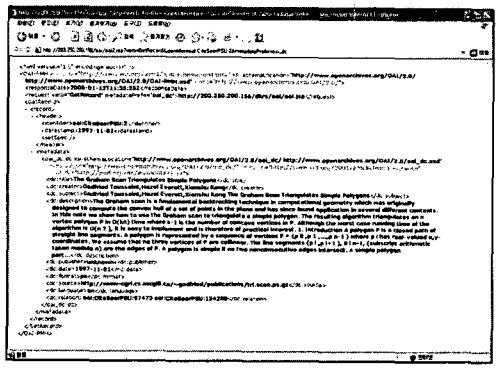

그림 8. GetRecord의 명령어 응답

5. 결론 및 향후 연구

OAI는 디지털 콘텐트의 상호운용성을 제공함으로써 효율적인 유통체제를 보급하는 역할을 수행한다. *stOAI* 시스템은 OAI 프로토콜을 통해 과학기술정보를 수집 · 저장하여 배포하는 시스템으로 그림 9와 같이 현재 해외 학술정보기관으로부터 약 60만 건의 데이터를 수집하여 제공하고 있다.

또한, 과학기술중앙정보센터의 역할을 수행하고 있는 KSITI의 과학기술관련 정보를 OAI 프로토콜을 통해 효율적으로 외부 정보 서비스연계 기관에 제공함으로써 학술정보 서비스의 유통체계에 큰 역할을 수행하고 있다.

향후, 수집 시스템을 통한 학술정보 데이터의 확대와 외부연계체제를 확충한다면 과학기술정보의 유통의 중요한 기틀로 자리 잡을 것이다.

참 고 문 헌

[1] Shreeves, S.L.;Kirkham, C.;Kaczmarek, J.;Cole, T.W.,"Utility of an OAI service provider search portal", Digital Libraries, 2003. Proceedings. 2003 Joint Conference on 27-31 May 2003 Page(s):306 - 308.

[2] Lagoze, C. and Van de Sompel, H."The Open Archives Initiative: Building a low-barrier interoperability framework" in Proceedings on ACM/IEEE Joint Conference on Digital Libraries (Roanoke VA, June 2001), ACM Press, 54-62.

[3] Carl Lagoze, Herbert Van de Sompel, Michael Nelson and Simeon Warner (editors) (2002), "The Open Archives Initiative Protocol for Metadata Harvesting", v2.0. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[4] The Distributed Library: OAI for Digital Library Aggregation related DLF projects: document authorship:, "The Case for OAI", last revised: 29 November 2005 Milewicz-v.8.

[5] 이상기, "OAI 기반 Open Digital Library 연구","A study on the OAI based Open Digital Library", 정보관리연구 2004-12-01 통권 제35권 3호.

[6]KERIS "생성유통체계", dCollection(국가지식정보생성 및 유통체계)시스템, http://www.dcollection.net