정보관리연구, vol.41, no.3 2010, pp.191-209

참조연계 서비스 구현을 위한 참고문헌 식별자 매칭 시스템

The Reference Identifier Matching System for Developing Reference Linking Service

이 용 식* · 이 상 기**
Yong-Sik Lee · Sang-Gi Lee

차 례

초 록

서로 다른 정보자원 간의 연결을 의미하는 참조연계 서비스를 위해서는 참고문헌 데이터베이스 구축과 식별자 매칭 작업이 필요하다. CrossRef, PubMed, Web Of Science 등의 많은 해외 기관들은 Inera의 eXstyles, Parity Computings의 Reference Extractor 등의 자동화 도구들을 이용하여 DOI, PMID 등의 식별자를 기반으로 하는 참조연계 체제를 구축하였다.

국내에서도 한국과학기술정보연구원, 한국연구재단 등의 여러 기관에서 참고문헌 데이터베이스를 구축하고 있다. 그러나 각 단체별로 채택하고 있는 다양한 참고문헌 기술 형식 때문에 망라적인 데이터베이스 구축은 많은 어려움에 직면해 있다.

이에 참고문헌을 자동으로 파싱하여 메타데이터를 추출하고, DOI, PMID, KOI의 식별자를 매칭하는 Citation Matcher 시스템을 개발하여 참고문헌 데이터베이스 구축의 효율성을 향상시키고자 한다.

키 워 드

참조연계, 참고문헌, 메타데이터, 식별자, Lookup, DOI, KOI, PMID

* 한국과학기술정보연구원 정보서비스실 연구원
(Researcher, Dept. of Information Service, KISTI, yslee92@gmail.com)
** 한국과학기술정보연구원 정보서비스실 선임연구원(교신저자)
(Corresponding Author, Senior Researcher, Dept. of Information Service, KISTI, sklee@kisti.re.kr)
• 논문접수일자: 2009년 8월 10일
• 최종심사일자: 2009년 9월 16일
• 게재확정일자: 2010년 5월 10일

ABSTRACT

A reference linking service that is connection of each other different information resource need to setup the reference database and to match identifier. CrossRef, PubMed and Web Of Science etc. the many overseas agencies developed reference linking service, that they used the automatic tools of Inera eXstyles, Parity Computings Reference Extractor etc. and setup in base DOI and PMID etc.

Domestic the various agencies of KISTI(Korea Institute Science and Technology of Information), KRF(Korea Research Foundation) etc are construction reference database. But each research communities adopts a various reference bibliography writing format.

As, the data base construction which is collect is confronting is many to being difficult.

In this paper, We developed the Citation Matcher System. This system is automatic parsing the reference string to metadata and matching DOI, PMID and KOI as Identifier.

It is improved the effectiveness of reference database setup.

KEYWORDS

Reference Linking, Reference, Metadata, Identifier, Identifier Lookup, DOI, KOI, PMID

1. 서 론

1.1 연구 배경 및 목적

해외의 주요 학술지 출판사들은 자체 출판 학술지의 일부 혹은 전체를 전자저널 형태로 웹에서 제공하고 있다. 전자저널이 제공하는 하이퍼링크의 기능은 연구자들이 학술지 논문을 읽으면서 관련된 다른 논문까지도 연속적인 참조를 가능하게 하고 있다(한혜영 외 2000).

과거의 디지털 환경에서는 전자자원 간의 연계를 위해서는 URL을 사용하였다. 그러나 URL 방식은 개별 사용자들이 위치정보를 지속적으로 유지 · 관리해야 하는 단점을 가지고 있다(Hoffma & Danial 1995).

이러한 URL의 단점을 개선하기 위해 IETF (Internet Engineering Task Force)에서 URN (Uniform Resource Name) 명세를 발표하였다. URN은 정보자원의 소장위치, 프로토콜, 호스트 등과는 상관없는 고유의 기호로써 영구적으로 식별되는 이름이다(Motas 1997).

URN은 그 특성상 어느 곳에서나 동일한 의미를 갖는 포괄성과 하나의 자원에는 하나의 URN만 부여되는 유일성을 가지고 있다. 또한 URN은 영구적으로 유일해야 하며 명명

기관이 존재하는 동안 참조정보로 사용될 수 있다(한혜영 외 2000).

이러한 URN의 특성을 활용하여 1997년에 미국출판협회와 CNRI(Corporation for National Research Initiative)에서 DOI(Digital Object Identifier)를 개발하였으며, 국내에서는 2004년 KISTI에서 KOI(Knowledge Object Identifier)를 개발하였다. DOI와 KOI는 모두 전자자원에 대한 식별자로서 사용할 수 있다.

참조연계를 위한 참고문헌 데이터베이스를 구축하기 위해서는 참고문헌을 제목, 저자, 학술지명, 발행년도, 권, 호 등의 메타데이터 단위로 파싱하고 각각의 인용정보에 대응하는 식별자를 저장하고 있어야 한다. 해외 주요 출판사들은 참고문헌을 분석하여 파싱하고 식별자를 검색하는 일련의 과정을 자동화한 Inera의 eXstyles, Reference Parity Computings의 Extractor 등과 같은 도구들을 활용하여 식별자 기반의 참조연계서비스를 구축하고 있다.

국내에서도 한국과학기술정보연구원(이하 KISTI)의 KSCI, 한국연구재단의 KCI, 한국의학회의 KoMCI 등에서 국내에 출판된 학술지를 대상으로 참고문헌 데이터베이스 구축 작업을 진행하고 있다. 하지만 국내 학술지의 참고문헌 서지 기술 규정을 적용하여 메타데이터를 추출하고 국내 논문에 대한 식별자 매칭이 가능한 자동화 도구는 전무한 상황이며 참고문헌 데이터베이스 구축 작업은 많은 인력의 투입을 통한 수작으로 이루어지고 있는 실정이다. 따라서 국내 학술지 참고문헌 환경에 적합한 자동화 도구의 개발은 참고문헌 데이터베이스 구축 작업의 효율성을 향상시키는 데 있어 반드시 필요한 상황이다.

본 연구에서는 참고문헌을 파싱하고 식별자를 검색하는 자동화 도구인 참고문헌 식별자 매칭 시스템(이하 Citation Matcher)을 개발하였다. Citation Matcher 시스템은 참고문헌을 제목, 저자, 학술지명, 권/호, 발행년도, 페이지 단위로 자동 파싱하고, 식별자로서는 DOI, PMID를 매칭하고 있으며 국내 논문에 대한 연계를 위해 KOI 식별자도 함께 매칭하고 있다.

Citation Matcher 시스템을 통해 대량의 참고문헌 데이터를 파싱하고 식별자를 매칭하며 이들 데이터를 시스템에서 자동으로 데이터베이스화하여 참고문헌 데이터베이스 구축 작업의 속도 및 효율성을 향상시킬 수 있다.

1.2 연구의 방법 및 범위

본 연구에서는 시스템의 설계를 위해서 KISTI에서 구축하는 KSCI의 참고문헌 데이터를 활용하였다. KSCI 데이터를 통해 권/호, 페이지, 발행년도, 학술지명 등의 각 항목별로 참고문헌 기술 형태를 분석하였다.

학술지 참고문헌에는 학술지 논문, 학위논문, 단행본, 연구보고서, 웹 자원 등의 다양한 형식의 서지사항이 기술되어 있다. 본 연구에

서는 참고문헌에 기술된 여러 형식 중에서 학술지 논문만을 대상으로 데이터를 파싱하고 식별자를 매칭할 수 있도록 연구하였다.

2. 관련연구

2.1 참조연계

참조연계는 하나의 정보자원과 다른 정보자원 간의 연결을 의미하는 일반적인 용어이다.

정보 간의 연계는 출판된 학술정보에 사용된 인용정보를 포함하여, 서지의 참고문헌 그리고 e-mail 등으로 전송되는 비형식적인 인용 등 다양한 방식으로 나타나고 있다. 최근까지는 학술지들 간의 참조연계가 개발되어지고 있으며, 점차 학술지를 넘어서는 연구들이 진행되고 있다.

과거에는 정보자원 간의 연결을 위한 도구로서 URL을 사용하였다. 하지만 URL이 가지는 문제점으로 인해 URL은 영구 식별자의 역할을 할 수 없다. 참조연계 시스템에서는 영구 식별자의 역할로서 IETF(Internet Engineering Task Force)의 URN 체계를 활용하고 있다.

URN은 시스템이나 프로토콜 등에 영향을 받지 않고 어느 곳에서나 동일한 의미를 가지며 있으며 하나의 자원에는 하나의 URN이 부여되는 유일성을 가지고 있다(Paskin 1999).

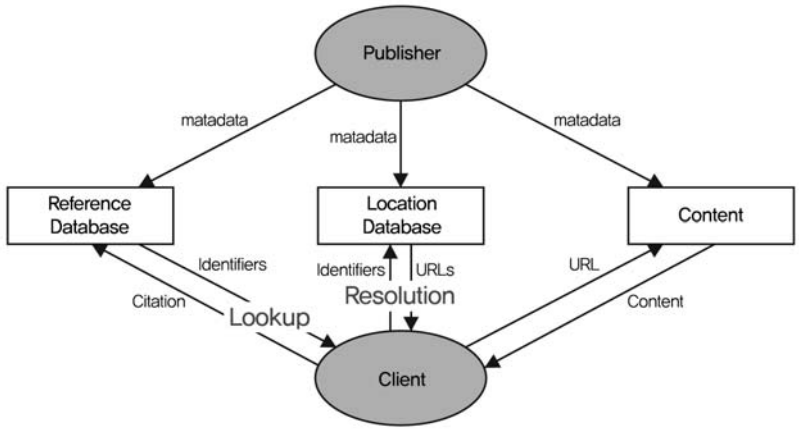
URN의 명세에 따라 구현된 대표적인 식별자로는 IDF(International DOI Foundation)의 DOI가 있다.

참조연계는 연계를 위한 데이터를 구축하는 방식에 따라서 정적 연계(static linking)와 동적 연계(dynamic linking)로 구분할 수 있다(Herbert Van de Sompel 1999).

정적 연계는 개개의 정보들 간의 연결 방법을 배치프로세스 방식으로 미리 계산하여 저장해 놓는 방식이다. 정적 연계의 사례로는 ISI의 Web Of Science와 NLM(National Library of Medicine)의 PubMed, CrossRef가 있다.

동적 연계란 연계가 필요한 시점에서 시스템에서 자동으로 연계를 제공하는 방식이다. 이 방식은 식별자를 필요한 시점에 계산해야 하므로 원문 위치와 식별자 정보 간의 불일치로 인한 연계 누락이 발생할 수 있다. 동적 연계의 예로는 영국전자도서관의 The Open Journal Project, Ghent 대학의 SFX, Openly Informatics의 S-Link-S(The Scholarly Link Specification)가 있다(한혜영 외 2000).

〈그림 1〉은 참조연계 시스템의 일반적인 모델이다. 각각의 정보자원들은 식별자를 가지고 있으며 개별 URL 형태로 여러 복사본이 존재하고 있다. 또한 하나의 정보자원은 각각의 식별체계별로 하나씩의 식별자를 가질 수 있다. 하나의 참조연계 시스템을 구성하기 위해서는 인용정보와 식별자를 저장하고 있는 참고문헌 데이터베이스(reference database), 식별자와 URL을 저장하고 있는 위치 데이터



〈그림 1〉 일반적인 참조연계 시스템 모델(Priscilla Caplan 1999)

베이스(location database)가 필요하다. 사용자들은 인용정보를 통해서 식별자를 검색하고 식별자를 통해서 URL을 검색하여 원하는 전자자원에 연계할 수 있다.

1) 참고문헌 데이터베이스(Reference Database)

각각의 정보에 있어서, 참고문헌 데이터베이스는 인용정보와 일치시킬 수 있는 최소한의 메타데이터와 식별자를 저장하고 있다. 하나의 콘텐츠에 관련된 인용정보를 찾기 원하는 사용자가 검색 질의를 전송하면 데이터베이스에서는 질의에 해당하는 식별자들의 목록을 반환한다. 이 과정을 일반적으로 'Reference lookup'이라고 한다.

Reference lookup을 위해서 참고문헌 데이터베이스는 제목, 저자, 학술지명 또는 ISSN, 발행년도, 권/호/편차 정보, 위치정보(페이지 또는 논문 번호), 매체 유형(논문, 초록 등의 구분) 등의 최소한의 메타데이터를 가지고 있어야 한다(Priscilla Caplan 1999).

2) 위치 데이터베이스(Location Database)

일반적으로 하나의 정보자원들은 여러 개의 복사본들이 존재하며 개별 URL의 형태로 웹에 위치하고 있다. 사용자는 식별자를 위치 데이터베이스에 전송하면 하나 이상의 URL을 전달받는다. 사용자는 URL들 중에서 하나의 URL을 선택한다. 식별자를 통해서 원문 URL을 획득하는 일련의 과정을 'resolution'이라고 한다.

CrossRef나 PubMed 등에서는 식별자를 통해 URL을 resolution 할 수 있는 API를 제공하고 있으므로 이들을 위치 데이터베이스로 활용할 수 있다.

2.2 식별자

참조연계를 위한 식별자는 반드시 아래의 세 가지 기능적 특성을 만족해야만 한다(Pri-

scilla Caplan 1999).

첫째는 지속성(persistence)이다. 식별자는 반드시 항구적으로 유지되어야만 하며, 식별자 서비스 기관은 기술적, 조직적으로 지속성을 충분히 보장해야 한다. PubMed의 PMID 같이 특별한 형식은 없으나 시스템적으로 잘 관리되어지는 개별 시스템의 키 값들도 식별자가 될 수 있다. 하지만 관리되지 않는 시스템에서 제공되는 키 값들은 식별자의 범위에서 제외된다.

둘째는 유일성(Uniqueness)이다. 식별자는 자신의 관리 범위 내에서는 반드시 유일한 값을 보장해야만 한다.

셋째는 다중 접근성(Multiple Resolution)이다. 하나의 정보자원은 웹상에서 개별 URL을 가진 여러 개의 복사본을 가질 수 있다. 따라서 하나의 식별자를 검색할 경우 사용자가 접근할 수 있는 모든 URL들을 반환해야 한다.

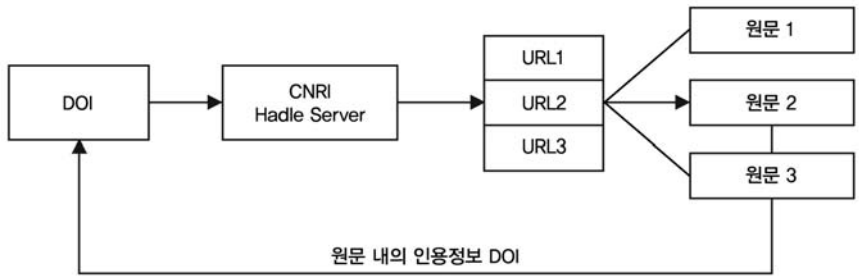
첫째와 둘째는 URN의 일반적인 특성이며 마지막은 참조연계를 위한 기능적인 특성이다. DOI와 PubMed의 ID인 PMID 등은 이 세 가지 조건을 만족하는 식별자들이다.

2.2.1 DOI

기존의 물리적 저작물이 디지털 저작물로 급속히 전환되면서 기존의 식별체계로는 디지털 콘텐츠의 특성을 충족시키기는 미흡하게 되었다. 디지털 콘텐츠의 특성상 각각의 오브젝트 즉 e-Book에서는 각 장별이나 페이지별, 그림이나 표 등을 따로 식별할 필요성이 있었으나 기존의 식별체계인 ISBN은 물리적 도서 한 권에 대한 식별 밖에는 하지 못하고 있다. 이는 콘텐츠 유통에서 매우 중요한 의미를 가지며 이와 같은 디지털 환경을 충족하기 위해 개발된 것이 DOI이다(안계성 2001).

DOI는 URN 명세를 충족시키면서 디지털 콘텐츠를 식별할 수 있는 식별체계로서 1997년에 미국출판협회와 CNRI에 의해 처음 개발되었다. 구문구조(Syntax)를 가진 식별체계로는 DOI가 가장 먼저 개발되었으며 2000년 5월에는 미국 국가 표준인 ANSI/NISO Z39.84-200으로 제정되었다(ANSI/NISO Z39.84-2000 2000).

DOI는 일정한 "/"로 구분되는 접두부와 접



〈그림 2〉 원문 내의 DOI 서비스 절차

미 형식의 구문 구조를 가진 식별자이지만 그 자체로는 데이터에 대한 특성을 알기 어려운 일종의 'dummy' 식별자라고 할 수 있다. 'dummy' 식별자는 URL을 연결하기 위한 URL 검색 및 변환 시스템을 필요로 하며, DOI에서는 CNRI Handle System을 사용하여 URL로 변환하고 있다(Miriam 2002, 〈그림 2〉 참조).

2.2.2 KOI

KISTI에서는 자관이 보유하고 있는 디지털 콘텐츠에 유일한 식별기호를 부여함으로써 사용자에게 식별기호만으로도 원하는 디지털 콘텐츠를 서비스 받을 수 있도록 KISTI 고유 식별기호인 KOI(Knowledge Object Identifier)를 개발하였다(이상환 2004).

KOI의 구문구조는 DOI, JOI와의 호환성을 고려하여 접두부와 접미부로 구성되어 있으며, 접두부와 접미부는 "/" 기호로써 구분된다.

KOI 역시 DOI와 동일하게 'dummy' 식별자이며 검색 및 변환 시스템으로서 RA 서버를 구축하여 KOI를 URL로 변환하고 있다.

KOI는 국내 학술논문, 학술회의, 연구보고서 등의 KISTI 보유의 국내 데이터에 한하여 부여되는 식별자이다.

2.2.3 PMID

PMID(PubMed IDentifier)는 PubMed에서 생명과학 관련 학술지 등에 인용된 PMC 자료들에 부여한 일련번호이다. 2005년까지 약 160만 개의 ID가 부여되었으며 매년 약 10만개의 데이터가 추가되고 있다. PMID는 일련번호로서 PMC에서만 사용 가능한 Local Key에 해당하지만 PubMed 시스템을 통해서 잘 관리되고 있으며 참조연계를 위한 식별자의 기능 세 가지를 모두 만족하므로 식별자라고 할 수 있다.

2.3 학술지 참고문헌 기술 형식

국내외의 주요 연구단체는 기관적 특성과 학문의 특성에 따라 고유하고 다양한 형태의 참고문헌 서지기술 방법을 사용하고 있다. 단체별로 생산되는 참고문헌의 다양성 때문에 망라적인 참고문헌 데이터베이스 구축은 현실적으로 많은 어려움에 직면하고 있다(남영준 외 2000).

해외의 주요한 참고문헌 서지 기술 규정을 살펴보면 미국의 3대 서지 기술 규정으로 APA 형식, MAL 형식, Chicago 매뉴얼이 있다.

APA 형식은 미국심리학회에서 제안한 참고문헌 서지기술 규정으로 이 규정은 심리학을 포함한 유관 기관인 행동과학, 간호학, 경영학, 경제학, 범죄학분야 등의 학술단체에서 주로 사용하고 있다. MLA 형식은 미국현대어문협회에서 제안한 참고문헌 서지기술 규정으로 125개 이상의 학술지 수전의 정기간행물, 잡지, 뉴스레터를 비롯하여 많은 대학출판물에서 사용되고 있다. Chicago 매뉴얼은 시카고 대학 출판부에서 제안한 참고문헌 서지기술 규정이다. 이 규정은 새로운 정보매체와 컴퓨터 조판기술의 발전 등 외부에 변화에 따라

〈표 1〉 국내외 참고문헌 기술 형식 규정

구분		참고문헌 기술형식
해외	APA형식	저자명. (출판년도). 논문제목. 학술지명. 권(호). 페이지.
	MLA형식	저자명. “논문제목.” 학술지명. 권. 호. (출판년도): 페이지.
	Chicago 매뉴얼	저자명. 출판년도. 논문제목. 학술지명. 권(호): 페이지.
	Turabian	저자명. “논문제목.” 학술지명 권, 호(출판월 년): 페이지.
	ISO R690	저자명. 논문제목. 학술지명. 출판년도, 권, 호, p. 페이지.
국내	한국심리학회	저자명 (출판년도). 논문제목. 학술지명. 권, 호, 페이지.
	한국정보관리학회	저자명. 출판년도. 논문제목.『학술지명』. 권(호):페이지
	한국정보과학회	저자명, “논문제목”, 학술지명, 권, 호, pp.페이지, 출판년도
	도서관정보학회지	저자명. “논문제목”, 학술지명, 권, 호 (출판년도), pp.페이지

상대적으로 빈번하게 개정되었다(남영준 외 2005).

〈표 1〉에서는 미국 3대 서지기술 형식 규정을 포함한 5개의 해외의 규정과 일부 국내 학회의 서지기술 형식 규정을 정리하였다. 각각의 학회 및 서지기술 규정은 인용부호나 각 항목의 순서, 권/호 표시 방식 등 거의 모든 부분에서 서로 다른 형식을 가지고 있음을 알 수 있다.

3. 참고문헌 서지 분석

3.1 권/호

〈표 2〉의 (1), (2)를 보면 같은 논문에 대해 참고문헌을 기술하는 경우에도 서지기술규정별로 ‘권/호’에 대한 표시가 서로 다름을 알 수 있다.

〈표 2〉의 (3), (4)는 권/호 정보에 관련된

〈표 2〉 참고문헌에서의 권/호 기술의 예

(1) 서승덕 (1965). ‘폭우의 시간적분포에 관한 고찰’, 한국농공학회지, Vol. 7, No.2, pp.792-7976
(2) 서승덕 (1965). ‘폭우의 시간적분포에 관한 고찰’ 한국농공학회지, 한국농공학회, 제 7권, 제 2호, pp. 792-797
(3) 권기록, 조아라, 이선구. 정맥주입용 산양산삼 증류약침의 급성, 아급성 독성 실험 및 Sarcoma-180 항암효과에 관한 실험적 연구. 대한약침학회지. 2003; 6(2): 7-27
(4) Palmer, M., Passonneau, R., Weir, C. & Finin, T (1993). The KERNEL text understanding system. Artificial Intelligence, 63, 17-68

명칭 표시 자체를 사용하지 않는 방법이며 특히 (4)의 경우는 '호' 부분이 누락되어 있다. 참고문헌 기술에 있어서 이처럼 특정 항목들이 누락한 경우가 상당수 존재하고 있다.

〈표 3〉은 권/호 패턴 표준을 위해 2004년 KISTI에서 권/호의 의미로 사용되고 있는 명칭들을 조사한 도표이다.

권/호에 대한 실제 명칭의 다양성은 시스템 설계를 어렵게 하는 요인이다. 참고문헌 파싱에 앞서 이들 명칭은 하나의 표준화 된(본 연구에서는 vol., no.을 사용한다) 명칭으로 변환해야 한다. 또한 권/호 이외의 pt., suppl. 과 같은 편차 정보들이 기술되어 있는 경우가 있으나 이들 편차 정보는 식별자 매칭에 사용되지 않는 부가 정보이므로 참고문헌 문자열에서 삭제한 후 파싱을 실행한다.

3.2 페이지

〈표 4〉는 페이지에 대한 다양한 기술 형식을 보여준다. 페이지 기호로서 사용되는 명칭을 살펴보면 'pp., p., page, PP., P., Page, 페이지, 쪽' 등 다양한 명칭이 있으며 〈표 4〉의 (3)에서 보듯이 페이지 표시를 생략하고 기술하는 경우도 있다. 이들 페이지의 명칭 역시 참고문헌 파싱 전에 하나의 표준화된(본 연구에서는 pp.을 사용한다.) 명칭으로 변환해야 한다.

〈표 3〉 권/호 명칭 매핑 테이블(KISTI 2004)

표준안	권/호 명칭
v.	권, 제권, volume, Volume, VOLUME, volumes, Volumes, VOLUMES, v, V, vol, Vol, VOL, vols, Vols, VOLS, Band, Bd., BAND, Jahrgang, Jahrg., t, tom, TOM, tome, tomo
no.	호, 제호, number, Number, NUMBER, numbers, Numbers, NUMBERS, n, N, no, No, NO, nos, Nos, NOS, num, Num, fascicle, Fascicle, fasc, Fasc, Numberg, numero, Numero, Heft, nr, Nr, nummer, Nummer, issue, issue no, issues
pt.	편, 제편, 파트, part, Part, PART, parts, Parts, PARTS, p, P, pt, Pt, PT, pts, Pts, PTS, parte, Parte, PARTE, partie
suppl.	부록, 증간, 보유, 보유판, 증보, supplement, Supplement, SUPPLEMENT, suppl, Suppl, SUPPL,, supplement no., supplement number, suppl. no., suppl number, appendix,

〈표 4〉 참고문헌에서의 페이지 정보 기술의 예

(1) J. M. Chang and N. F. Maxemchuk, 'Reliable Broadcast Protocol,' ACM Transactions on Computer Systems 2(3), 1984, pp.251-273
(2) Pavel V, Weidinger SBK and Kovarik P (2000) Distraction displays in meadow pipit (Anthus pratensis) females in central and northern Europe. Ethology 106: 1007-1019
(3) 이지홍, 전봉환, '4족 보행 로봇의 동적 조작도 해석,' 대한전자공학회 2003 하계종합학 술대회 논문집, 2721-2724쪽, 2003년 7월

3.3 날짜

대부분의 참고문헌에서는 출판년도 만을 표시하고 있으며 출판년도의 위치나 ()의 사용 여부 등에서만 차이가 있다. 날짜 표시 형식에서 분석이 필요한 부분은 'Turabian' 서지기술 규정(〈표 5〉 참조)과 같이 출판년도와 함께 출판월을 같이 입력하도록 하고 있는 경우이다.

출판월은 표준화된 입력 방식이 없으며 저자별로 입력하는 방식이 다양하다고 할 수 있다. 〈표 6〉은 참고문헌 내의 월별 명칭을 분석한 것이다. 출판월은 참고문헌 식별자 검색에 사용되지 않는 부분이므로 본 연구에서는 출판월을 제거한 후 참고문헌 문자열의 파싱을 실행한다.

〈표 5〉 참고문헌에서의 날짜 기술의 예

(1) A. Zelinsky, 'A mobile robot exploration algorithm,' IEEE Transactions on Robotics and Automation, vol. 8, no. 6, pp. 707-717, Dec. 1992
(2) K. R. Williams and R. S. Muller, 'Etch Rate for Micromachining Processing, Journal of Microelectro-mechanical Systems,' Vol.5, No.4, December, pp.256-269, 1996

〈표 6〉 참고문헌 내의 월별 명칭

월	표 기 방 법
1월	January, JANUARY, January, jan, jan., Jan., 1월
2월	February, FEBRUARY, February, feb., Feb., FEB., 2월
3월	March, march, mar., Mar., MARCH, 3월
4월	April, April, Apr., APRIL, apr., APR., 4월
5월	May, may, MAY, 5월
6월	June, Jun., jun., JUNE, JUN., June, 6월
7월	July., JULY., July., jul., Jul., JUL., 7월
8월	August, AUGUST, August., Aug., aug., AUG., 8월
9월	September, september, SEPTEMBER, sept., Sept., SEPT., sep., Sep., SEP., 9월
10월	October, OCTOBER, October., oct., Oct., OCT., 10월
11월	November, November, NOVEMBER, Nov., nov., NOV., 11월
12월	December, DECEMBER, december, Dec., dec., DEC., 12월

3.4 학술지명

참고문헌에 학술지명을 기술함에 있어서 다양한 형태의 약어들이 사용되고 있다. 학술지명에 대한 표준 약어명이 존재하지 않기 때문에 같은 학술지에 대해서도 저자별로 다양한 형태의 약어로 기술되고 있다(〈표 7〉 참조).

〈표 8〉은 참고문헌에서 사용되는 대표적인 약어들에 대해 정리하였다. 학술지명에 사용되는 약어는 그 형태가 매우 다양하여 모든 약어들을 정식 명칭으로 변환하는 것은 대단히 어려운 작업이다. 본 연구에서는 〈표 4〉에 정리된 대표적인 약어들만을 정식 명칭으로 변환한 후 참고문헌 파싱을 실행한다.

4. Citation Matcher 시스템 설계 및 구현

Citation Matcher 시스템은 참고문헌이나 메타데이터를 입력 받아서 메타데이터를 분리하고 DOI, PMID, KOI 등의 식별자를 반환하는 시스템이다(〈그림 3〉 참조). 저자가 직접 입력한 참고문헌의 경우 데이터가 부족하거나 부정확한 경우가 많다. Citation Matcher에서는 DOI, PMID, KOI 식별자를 매칭하고 부정확한 메타데이터를 정제하여 반환한다.

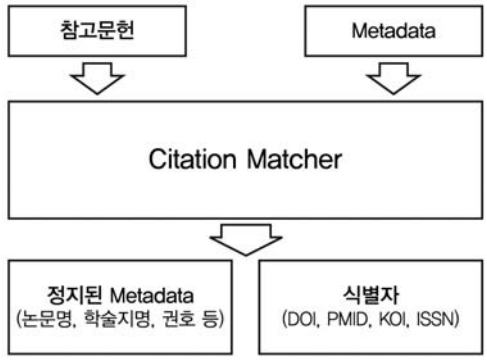
Citation Matcher 시스템은 정규화(Normalizer), 메타데이터 분리(Parser), 식별자매칭(Matcher) 세 가지 모듈로 분리하여 설계하였다(〈그림 4〉 참조).

〈표 7〉 참고문헌에서의 학술지명 기술의 예

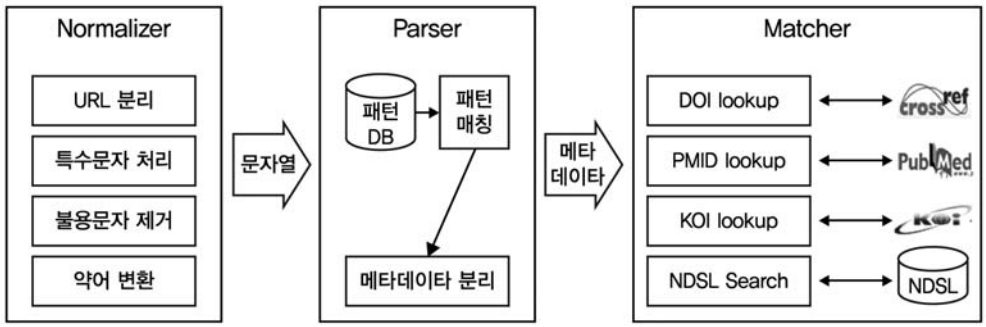
(1) Farzin Mokhatarian and Riku Suomela, 'Robust Image Corner Detection Through Curvature Scale Space,', *IEEE Tr. on PAMI,* VOL. 12, December 1998
(2) F.H. Harlow and A.A. Amsden, 'A numerical fluid dynamics calculation method for all flow speeds,' *J. of Comp. Phy.,* 8 197(1971)
(3) 정형환, 정문규 , 유재엽, '퍼지 전 보상 PID 제어기를 이용한 전력계통의 부하주파수 제 에 관한 연구', *生産技術硏究所論文集,* 제 4권 1호, pp.241-248, 1999

〈표 8〉 참고문헌 내의 약어 사용 예

정식 명칭	약 어
Journal	J., Joun., Jour., Journ.
Proceedings	proc. Proc. Procd., Pro., pro.
Conference	conf., Conf., Con.,
International	int. Int., Inter., inter., Int'l
Transaction	Trans., trans., Tr., tr., Tran., tran., Tran., trans.
Letter	Let. Lett., let., lett.
Physics	Phy., PH., Phys., phys. PHys.



〈그림 3〉 Citation Matcher 시스템 개요



〈그림 4〉 Citation Matcher 시스템 구성

정규화는 참고문헌에 대한 분석 내용을 반영하여 참고문헌 문자열의 정규화 과정을 실행한다. 메타데이터분리는 정규식을 이용한 패턴매칭 기법을 활용하여 참고문헌을 메타데이터 단위로 분리한다. 식별자매칭에서는 CrossRef, PubMed, KOI에서 제공하는 lookup API를 활용하여 식별자를 검색하며, KISTI의 NDSL 데이터베이스를 매칭 알고리즘에 활용하였다.

4.1 정규화(Normalizer)

참고문헌 분석에 의하면 논문의 참고문헌에는 권, 호, 페이지, 인용 등의 표시에 있어서 다양한 형태의 명칭을 사용하고 있다. 또한 학술지명의 표기에서는 통일되지 않은 약어의 사용이 빈번하다. 참고문헌의 식별자 매칭의 효율을 향상시키기 위해서는 이러한 이형들에 대한 표준화 및 불필요한 데이터 제거 등의 정규화 과정이 필요하다.

이에 본 연구에서는 CiteSeer의 Automatic Citation Indexing System의 파싱 알고리즘[1]과 참고문헌 데이터의 분석을 통해 아래와 같은 정규화 알고리즘을 정의하였다.

① 참고문헌에 포함된 URL은 자체로 링크

에 기능을 가지고 있으나 식별자 검색을 위한 메타데이터에는 포함되지 않기 때문에 문자열에서 분리한다.

② 전체 문자열을 소문자로 변환한다.

③ 참고문헌에는 여러 가지의 특수 문자들이 사용되고 있다. 그 중에서 Ⅰ,Ⅱ,Ⅲ,Ⅳ,Ⅴ 등의 로마 숫자는 아라비아 숫자로 변환하고, 『 』, “ ” 와 같은 인용 표시는 ASCII 문자인 "로 변환한다.

④ 참고문헌에는 메타데이터와 관계가 없는 불용 문자나 단어들을 포함되어 있다.

대표적으로 참고문헌 첫머리의 (1), [1], 1. 과 같은 인덱스 표시와 et al, etc 등의 의미 없는 불용 단어는 제거한다. 또한 날짜 표시에 사용되는 월 관련 약어들(〈표 6〉 참조)은 식별자 검색에 사용되지 않기 때문에 파싱의 효율 향상을 위해서 사전에 모두 삭제한다. 또한 권/호/페이지의 명칭으로 사용된 단어들은 모두 vol., no., pp.으로 치환하면 불필요한 편차 정보들은 모두 삭제한다.

⑤ 학술지명에 사용되는 많은 약어들 중에 대표적인 약어들은 정식 단어로 변환시킨다. 예로써 conf.는 conference로 proc.는 proceedings로 확장한다.

4.2 메타데이터 분리(Parser)

참고문헌의 파싱은 정규식을 활용한 패턴 매칭 기법을 사용하였다. 참고문헌 패턴을 메타데이터 구성 순서와 권/호/페이지 명칭의 사용 여부, 데이터 누락 여부, 제목에 인용 기호 사용여부에 따라서 패턴을 세분하여 문자열의 패턴 매칭을 실행하고 각각의 세부 패턴별로 독립적인 파서 프로그램을 작성하였다. 현재 62개의 정규식 패턴을 작성하여 파싱에 활용하고 있다.

〈표 9〉는 파서에서 사용하고 있는 정규식 패턴을 구성하는 방식에 대한 예제이다. 세부

〈표 9〉 참고문헌 패턴

패 턴 명	패턴 구성
APA_1	저자명, 발행년도, ‘제목’, 학술지, vol.권, no. 호, pp. 페이지
APA_2	저자명, 발행년도, ‘제목’, 학술지, vol. 권, pp 페이지
APA-3	저자명, 발행년도, 제목, 학술지, vol. 권, pp 페이지
정보과학회_1	저자명, ‘제목’, 학술지, vol. 권, no. 호, pp 페이지, 발행년도
정보과학회_2	저자명, ‘제목’, 학술지, Vol. 권, pp 페이지, 발행년도
정보과학회_3	저자명, ‘제목’, 학술지, 권, 페이지, 발행년도
정보과학회_4	저자명, ‘제목’, 학술지, vol. 권, no. 호, 발행년도
정보과학회_5	저자명, ‘제목’, 학술지, vol. 권, no. 호, pp 페이지

1) CiteSeer에서는 다음 5 단계를 제시하였다.
① 소문자로 변환 ② ‘-’ 제거 ③ [3], (3), [Giles 92]와 같은 인용 인덱스 제거 ④ 약어의 확장 ⑤ et. al., &, (), {} 등과 같은 불필요한 단어 및 문자 제거

적인 패턴을 구성하여 파싱하는 방법을 사용하고 있어, 정확한 패턴의 구성과 많은 수의 참고문헌 패턴의 제작이 시스템 성능을 좌우한다.

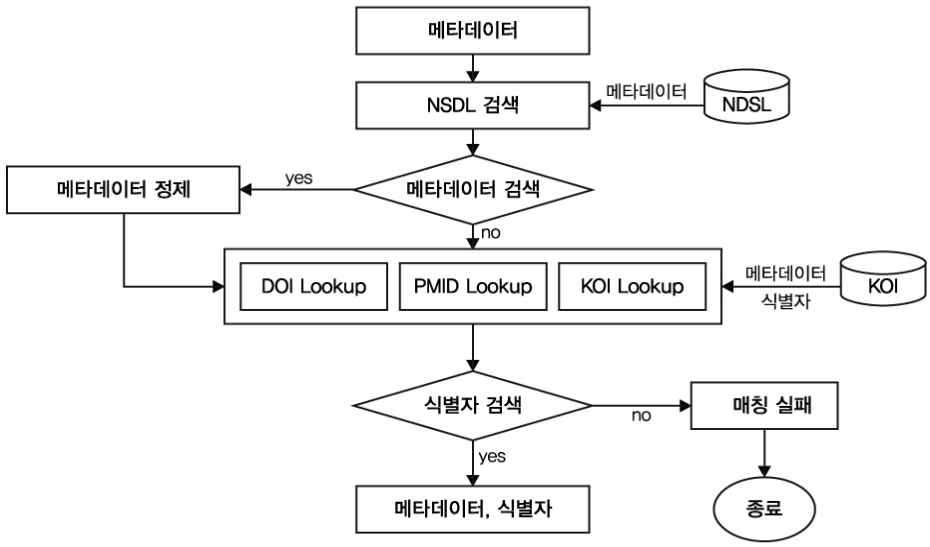
4.3 식별자 매칭(Matcher)

식별자의 매칭 작업은 기본적으로 Cross-Ref, PubMed에서 제공하는 식별자 Lookup API를 사용하고 KOI의 경우는 KISTI 제공 서비스로 직접 KOI 데이터베이스를 검색하여 식별자를 매칭하도록 알고리즘을 구성하였다.

또한 KISTI에서 서비스하는 NDSL에서는 약 5천만 건의 국내외 학술데이터를 소장하고 있으며, 이 데이터를 식별자 매칭 알고리즘에 활용하였다. NDSL 데이터베이스를 1차 검색하여 참고문헌 데이터를 정제하고 매칭의 성능을 향상시키고자 하였다.

〈그림 5〉의 식별자 매칭 프로세스를 살펴보면 다음과 같다.

파서를 통해서 분리된 메타데이터들은 먼저 NDSL 데이터베이스의 검색 과정을 거치게 된다. NDSL에서 검색된 메타데이터를 활용하여 저자가 입력한 데이터의 누락이나 오류 부분을 수정하는 데이터 정제 작업을 진행한다. 이후에는 CrossRef, PubMed에서 제공하는 식별자 Lookup API를 사용하여 식별자 검색 작업을 진행한다. KOI의 경우는 직접 KOI 데이터베이스에 접속하여 논문제목, 권, 호, 페이지, 발행년도의 메타데이터를 활용하여 식별자를 검색한다.



〈그림 5〉 매칭 프로세스 흐름도

4.4 참고문헌 매칭 서비스 구현

Citation Matcher 시스템은 현재 KISTI의 CLICK 시스템을 통해서 서비스되고 있다 (http://click.ndsl.kr). 웹을 통해서 Single Citation Matcher, Multi Citation Matcher, Free Citation Matcher, Batch Citation Matcher 네 가지 기능으로 나누어 서비스하고 있다. Single Citation Matcher, Multi Citation Matcher, Batch Citation Matcher 의 세 가지는 참고문헌에 사용되는 메타데이터를 모두 알고 있을 경우 사용하는 기능이다.

Free Citation Matcher의 경우는 논문의 참고문헌 텍스트를 copy & paste하여 사용한다. 참고문헌 텍스트를 자동으로 파싱하고 식별자 매칭 작업을 실행한다. 〈그림 6〉은 Free Citation Matcher를 통해서 참고문헌을 파싱하고 식별자 매칭을 실행한 결과 화면이다. '입력 서지정보'에서는 참고문헌 텍스트를 파싱한 내용을 보여주며 '매핑 서지정보'에서는 NDSL과, KOI 데이터베이스를 통해 정제한 메타데이터를 출력한다. '식별자' 부분에서는 매칭한 DOI, PMID, KOI와 같은 논문에 해당하는 식별자와 ISSN과 같은 학술지에 해당하는 식별자를 출력하고 있다.

〈그림 6〉 Free Citation Matcher 실행 결과 웹 화면

〈그림 7〉은 Citation Matcher의 실행 결과를 XML로 전송받은 화면이다. 사용자들은 XML로 전송받을 결과를 활용하여 참고문헌 데이터베이스 구축에 활용할 수 있다. 웹 화면을 통하지 않고 프로그램 상에서 프로토콜을 활용하여 URL을 통해 데이터를 입력받아서 Citation Matcher를 실행하여 그 결과를 XML 파일로 검색 결과를 제공하는 Web API를 제공하고 있으며 사용법은 아래와 같다.

http://click.ndsl.kr/servlet/getService.do?refStr=〈참고문헌&doiFlag=〈0,1〉&pmidFlag=〈0,1〉&koiFlag=〈0,1〉

- refStr : 참고문헌 문자열
- doiFlag : DOI 매칭 여부 (1 : 사용 0: 사용안함)
- koiFlag : KOI 매칭 여부 (1 : 사용 0: 사용안함)
- pmidFlag : PMID 매칭 여부 (1 : 사용 0: 사용안함)

이들 모듈을 활용하여 참고문헌 데이터베이스 구축 작업을 자동화 할 수 있다.

5. 분석 및 결론

KISTI의 KSCI에서 구축한 참고문헌 데이터베이스를 사용하여 시스템 테스트를 진행하였다. KSCI에는 현재 약 160만 건의 데이터가 구축되어 있으며 이들 중에서 학술지 논문

```
- <outputData>
  - <journal number="2">
    - <journalTitle ndslID="10121448">
        <![CDATA[ infect immun ]]>
      </journalTitle>
      <pissn>0019-9567</pissn>
      <eissn>1098-5522</eissn>
      <publisher />
      <year>2001</year>
      <volume>69</volume>
      <issue />
    </journal>
  - <article number="2">
    - <articleTitle>
        <![CDATA[ Invasion of Human Epithelial Cells by Pseudomonas aeruginosa Involves Src-Like Tyrosine Kinases
        p60Src and p59Fyn ]]>
      </articleTitle>
    - <authors>
        <author number="1">Esen, M.</author>
        <author number="2">Grassme, H.</author>
        <author number="3">Riethmuller, J.</author>
        <author number="4">Riehle, A.</author>
        <author number="5">Fassbender, K.</author>
        <author number="6">Gulbins, E.</author>
      </authors>
      <startPage>281</startPage>
    - <identifiers>
        <doi>10.1128/IAI.69.1.281-287.2001</doi>
        <url />
        <koi />
        <nid>10121448</nid>
        <pmid />
      </identifiers>
    </article>
  </outputData>
</record>
```

〈그림 7〉 Citation Matcher 실행 결과 XML

에 해당하며 DOI, KOI가 식별된 데이터는 19만5,499건(2009.08.10 기준)이 있다.

본 연구에서는 이들 19만 건의 데이터 중에서 3만 건을 임의로 추출하여 매칭 테스트를 진행하였다.

〈표 10〉은 KSCI에 대한 현황 분석 데이터이다. 19만 건의 데이터 중에서 DOI가 매칭된 데이터는 약 11만 건, KOI는 약 8만 건의 데이터가 있다.

〈표 11〉의 Citation Matcher를 실행한 결과 구축된 데이터에 대한 분석이다. 참고문헌에서 메타데이터를 추출하는 파싱 과정은 2만6,408건이 성공하여 83%의 성공률을 보이고 있으며 최종 매칭에 성공한 데이터는 2,630건으로 매칭 성공률은 65%로 측정되었다. 약 35% 정도의 데이터에 대한 식별자 매칭에 실패하였다.

〈표 10〉과 〈표 12〉를 비교했을 때, DOI에 대한 매칭의 비율은 차이가 없는 반면에 KOI에 대한 매칭에 있어서 많은 누락이 발생했음을 알 수 있다. 이는 KOI 식별자를 매칭함에 있어서 시스템에서는 참고문헌 메타데이터와 KOI 데이터베이스의 일치 검색을 실행하고 있으나 실제 메타데이터는 권, 호, 페이지 정보의 누락이나 제목, 학술지명에서의 오타 등의 일치 검색만으로는 매칭이 불가능한 데이터가 많았기 때문으로 분석된다.

파싱에 있어서 약 20% 정도의 실패는 참고문헌에 대한 세부 패턴의 부족이 원인으로 분석되었다. 'W. J. Alford, J. Chem. Phys.

〈표 10〉 KSCI 매칭 데이터 분석(학술논문)

	전 체	DOI	KOI	PMID
건수	195,499	110,995	84,889	0
비율		57%	43%	0%

〈표 11〉 Citation Matcher 실행 결과 분석

	전 체	파싱 성공	매칭 성공
건수	31,708	26,408	20,630
비율		82%	65%

〈표 12〉 식별자별 실행 결과 분석

	DOI	PMID	KOI
매칭 건수	18,668	306	4,370
매칭 비율	59%	1%	13.8%

96, 4330, 1992'와 같이 학술지명이 없는 패턴과 같은 일부 데이터가 누락되어 있는 패턴들이 파싱에 실패하였다.

해외의 주요 사이트들은 하나의 논문을 검색하면서 참조된 다른 논문들까지 연속적으로 검색할 수 있는 참조연계 서비스를 구축하고 있다. 참조연계 서비스의 구축을 위해서는 참고문헌에서 메타데이터를 추출하고 DOI와 같은 식별자 매칭하여 참고문헌 데이터베이스를 구축하는 과정이 필수적이다.

국내에서도 한국과학기술정보연구원, 한국연구재단 등의 여러 기관에서 참고문헌 데이터베이스를 구축하고 있다. 그러나 참고문헌 기술 규정의 다양성과 권/호/페이지 등에서 많은 이형의 명칭을 사용하는 문제 등으로 인해 데이터베이스 구축 작업에 많은 어려움을 가지고 있다.

본 연구에서는 참고문헌을 파싱하여 메타데이터를 추출하고 식별자를 매칭하는 자동화 시스템을 연구하였다. Citation Matcher 시스템을 통하여 대량의 참고문헌 데이터에 대한 메타데이터의 추출과 식별자를 매칭, 자동 데이터베이스 구축 작업이 가능하게 되었으며, 시스템을 활용하여 참고문헌 데이터베이스 구축 작업의 속도 향상 및 효율성 개선이 가능하였다.

하지만 위의 분석 결과를 살펴보면 참고문헌 데이터베이스 구축 작업의 완전 자동화를 위해서는 파싱율과 KOI에 대한 매칭율의 향상이 필요함을 알 수 있다. 참고문헌에 대한 세부 패턴을 추가하여 파싱율을 높이고 KOI 매칭 프로세스를 개선하여 전반적으로 매칭율을 보다 향상시켜야 한다.

학술논문의 참고문헌의 인용 상황을 살펴보면 학술논문 외에도 단행본이나 연구보고서, 특허, 웹 자원에 대한 인용 서지가 기술되어 있다. 향후 시스템의 발전을 위해서는 현재 학술지 논문에만 한정되어 있는 시스템을 단행본과 프로시딩 등의 이종 자료들의 확장이 필요하다.

참고문헌

김홍렬, 정경희. 2005. 국내 참고문헌 데이터베이스 운영현황 및 실태에 관한 분석.『정보관리학회지』, 22(2): 23-39.

남영준, 조현양, 배순자. 2005. 참고문헌 서지기술 표준에 관한 연구.『한국문헌정보학회지』, 39(4): 261-279.

안계성. 2001. DOI/INDECS를 이용한 디지털 컨텐츠 보호.『정보보호학회지』, 11(5): 52-62.

이상환, 신동구, 김재수, 최진영, 정택영. 2004. 식별체계기반의 전자원문 연계시스템 설계 및 구현.『정보관리학회지』, 21(3): 15-29.

한혜영, 정동열. 2000. 국내 학술지 논문의 DOI 기반 연계시스템 구축에 관한 연구.『정보관리학회지』, 17(4): 207-227.

정택영 외. 2004.『과학기술잡지 권/호 패턴 표

준』. 대전 : 과학기술정보표준화위원회.

ANSI/NISO. 2000. "Syntax for the Digital Object Identifier", Z.39-84-2000.

C.L. Giles, K.D. Bollacker, S Lawrence. 1998. "CiteSeer: An automatic citation indexing system." *Proceedings of the third ACM conference on Digital libraries*, 89-98.

Herbert Van de Sompel. 1999. "Reference Linking in a Hybrid Library Environment. Part 1: Frameworks for Linking." *D-Lib Magazine*, 5(4).

Hoffman, Paul and Daniel, jr. 1995. "URN Resolution Overview". Internet Draft, October 1995, IETF URI Work Group. [cited 2009.08.05]. 〈http://ftp.ics.uci.edu/pub/ietf/uri/draft-ietf-uri-urn-res-descript-oo.txt〉.

Miriam E. Blake, Frances L. Knudson. 2002. "Metadata and reference linking." *Library Collections, Acquisitions, and Technical Services*, 26(3): 219-230.

Norman Paskin. 2003. The DOI® Handbook. International DOI Foundation. [cited 2009.08.01]. 〈http://www.doi.org/handbook_2000/DOIHandbook-v3-3.pdf〉.

Priscilla Caplan, William Y. Arms. 1999. "Reference Linking for Journal Articles." *D-Lib Magazine*, 5(7/8).

R.Motas. 1997. "URN Syntax", RFC2141. [cited 2008.12.28]. 〈http://www.ietf.org/rfc/rfc2141.txt〉.

CrossRef System specification. [cited 2009.08.01]. 〈http://doi.crossref.org/doc/SystemInterface.html〉.

Crossref batch query for DOI lookup [cited 2009.08.01]. 〈http://doi.crossref.org/doc/Query.html〉.