메타데이터를 이용한 '학술적 웹 디렉토리 서비스'

본고는 메타데이터를 이용하여 인터넷 정보자원에 대한 디렉토리 서비스를 제공하는 사례를 소개한 것으로서 '정보의 과학과 기술' 49권 1호(1999.1)에 실린 '메타데이터를 이용한 학술적 웹 디렉토리 서비스의 구축'을 완역한 것이다. 최근 국내에서 추진되고 있는 인터넷상의 특정 분야 메타데이터 데이터베이스 구축에 참고가 될 것으로 사료된다.

안현수/한국통신 연구개발본부

국제대학 글로벌 커뮤니케이션 센터에서는 웹 정보자원의 인터넷 디렉토리인 ORel(Online Resource Locator)을 웹상에서 구축, 운용하고 있다. 이 디렉토리는 학술 목적의 논문 등과 같이 웹상에 산재한 정보자원을 추출한 후, 그 정보자원에 대한 서지정보데이터(메타데이터)를 데이터베이스화하여 홈페이지 상에서 검색이 가능하도록 한 서비스이다. 본고는 ORel의 구축 과정을 메타데이터와 관련지어 소개하고 구축시 유의해야 할 점들에 대해 검토한다.

ORel의 추진과정과 목표

ORel 프로젝트는 국제대학 글로벌 커뮤니케이션 센터(약칭 글로컴) InfoJapan 프로젝트로서 1994년도에 시작되었으며 '일본에서의 우수한 정보제공'을 목적으로 하고 있다(이 프로젝트는 주식회사 아스키와의 연구협력을 통해 실현되었다). 프로젝트 개시 당시에는 사회과학 분야의 연구기관에서 인터넷에 접속하는 경우가 매우 드물었기 때문에 InfoJapan에서의 '일본에서의 정보제공은 정보제공을 대행하는 것'으로의 측면이 강했다.

인터넷이 보급되면서 정보제공 대행으로서의 목적 의미는 약해졌으며 1997년에는 이러한 방침을 전환하였다. 즉, 정보제공을 대행하는 것보다는 제공된 정보를 '선택'하도록 함으로써 인터넷 이용자에 대하여 일본에서 제공하는 우수한 정보에 대한 액세스를 제공하는 방향으로 목표를 잡았다.

인터넷은 최근 수년동안 급속히 일반에게 보급되었으며 연구자 개인의 정보제공 수단으로도 이용되고 있다. 연구자가 개인으로 개설한 홈페이지에는 논문 등과 같은 문헌 정보자원이 게재되는 것이 드물지 않게 되었다. 웹은 일반적으로 이용자에게 있어서도 연구업적을 공개하기 위한 대체 수단 중의 하나가 되고 있으나 이미 널리 인식되고 있는 바와 같이 웹에서 정보를 적절히 검색하는 것은 쉬운 일이 아니다.

어려운 이유 중의 하나는 인터넷이 많은 콘텐트가 실려있는 매체이기 때문에 목적에 맞는 정보들을 구별하여 검색하는 것이 곤란하기 때문이다. 잡다한 정보가 실려있기 때문에 의외로 새롭게 결합된 정보가 발견되는 경우도 있지만 어느 특정 영역의 정보만을 입수하고자 할 경우에는 역으로 약점을 갖는다.

이러한 약점을 부분적으로 극복하기 위해 국제대학 글로벌 커뮤니케이션 센터에서는 웹상에 공개된 학술문헌 중에 인문사회과학분야의 학술문헌을 대상으로 디렉토리를 작성한 후, 등록 문헌의 검색이 가능한 서비스를 제공하고 있다.

ORel이라고 불리우는 이 서비스는 웹상에서 제공되고 있다. 문헌 정보자원의 디렉토리검색서비스를 제공함으로써 웹을 통한 학술정보제공의 가능성을 넓혔으며 인터넷의 발전에 기여하는 것이 ORel 프로젝트의 최종 목표이다.

인터넷상의 학술문헌 검색서비스로는 일본학술정보센터 등과 같은 도서관 분야의 목록서비스와 디렉토리서비스가 널리 이용되고 있다. 이러한 다양한 서비스와 ORel은 인터넷상에서 디렉토리서비스를 제공한다는 점에서는 공통되지만 서비스 대상으로서의 정보자원이 다르다.

ORel 데이터베이스 구성

ORel에서는 문헌 정보자원의 저자명, 제목, URL 등과 같은 메타데이터들은 데이터베이스에 등록하지만 문헌정보자원 그 자체는 수록하지 않는다. 인터넷상에서 문헌정보자원을 모아놓는 사이트(BibEc 등)도 있지만 ORel은 어디까지나 정보원에의 액세스를 한번에 제공하기 위한 서비스이다. ORel에 등록된 문헌 정보자원의 본문을 입수하기 위해서는 URL을 통해 액세스하는 것이 가능하다.

ORel에서의 메타데이터 부여는 1997년 11월 시점의 더블린 코어의 스키마를 참고하였다. ORel의 목적은 온라인 문헌 정보자원을 실제로 수집하여 디렉토리로 제공하는 것이기 때문에 메타데이터를 위한 스키마는 범용성이 있는 기존의 것을 채용할 필요가 있다.

더블린 코어는 네트워크 객체, 특히 문서형 객체의 속성기술을 위해 제안되었기 때문에 일반적인 속성 항목군이며 ORel의 취지에서 볼 때에도 이용하기 쉬운 시스템이었다는 점이 더블린 코어를 채용한 이유들 중의 하나이다.

더블린 코어에는 다음과 같은 15개의 필드가 제안되어 있다.

- 제목(title)
- 저자 또는 작성자(creator)
- 주제 또는 키워드(subject)
- 개요(description)
- 발행자(publisher)
- 다른 관계자(contributor)
- 일자(date)
- 정보자원의 종별(type)
- 형식(format)
- 정보자원 식별자(identifier)
- 출전(source)
- 언어(language)
- 관계(relation)
- 범위(coverage)
- 권리처리(rights)

더블린 코어는 네트워크에서 제공되는 정보자원에 대한 記述의 스키마를 필요충분한 형태로 제시한 것은 아니다. 고도의 지식과 경험을 축적한 전문가에 의한 목록의 이점과 정보자원 제공자가 자유롭게 정보자원을 공개할 수 있는 인터넷의 이점 이 두가지를 살려서 전문가가 아닌 정보자원 제공자일지라도 메타데이터를 쉽게 제공할 수 있을 정도로 간략화된 스키마가 바로 더블린 코어인 것이다. 더블린 코어의 15개 필드는 메타데이터 기술에 있어서 최소한의 필드를 나타내며 필요에 따라 필드를 추가할 수 있다.

ORel의 메타데이터는 기본적으로 더블린 코어에서 제안된 필드를 기반으로 하고 있다. 다음에서는 ORel에 등록된 메타데이터의 항목을 나타낸다. 괄호안에는 대응되는 더블린 코어의 필드가 있다. 현시점에서는 관계(relation), 범위(coverage), 권리처리(rights) 3개의 필드는 사용되지 않고 있다.

명칭(DC.Title)

문헌정보자원의 명칭. 일본어와 영어.

저자(DC.Creator)

문헌정보자원의 저자. 복수로 존재하는 경우에는 복수를 열거. 일본어와 영어.

번역자(DC.Contributor)

문헌정보자원을 번역한 인물. 복수 존재하는 경우에는 복수를 열거.

기술(記述)언어(DC.Language)

문헌정보자원이 기술된 언어. 번역 정보자원의 경우에는 번역판에 사용된 언어

연구영역(DC.Subject)

ORel에서의 정보자원의 대분류를 위해 사용된다. 이 필드는 반드시 전통적 학문 분류에 따르지 않고 Orel에서 독자적으로 할당한다.

주제(DC.Subject)

문헌정보자원이 다루는 주제를 위한 필드. 이 필드는 만약 사회과학분야일 경우의 Social Sciences Index 등과 같이 그 문헌정보자원이 관계하는 분야의 표준적인 어휘를 이용하여 기술하고 있다.

키워드(DC.Subject)

정보자원 중의 주요 어구 또는 개념

개요(DC.Description)

문헌정보자원의 요약 또는 개요

URL(DC.Identifier)

인터넷상의 식별자인 URL을 위한 필드

종별(種別)

문헌의 종류를 표현하기 위한 필드. 단독의 문서, 전자 저널, 보고서의 온라인 판의 별쇄 등이 이 필드를 이용하여 표현된다.

데이터 형식

온라인화에 이용되는 데이터형식. 주요 형식으로는 HTML, PDF, 일반 텍스트 등.

공개년월일

본 정보자원이 공개된 연월일.

출전(出典)

본 정보자원이 최초로 공개된 출판매체(해당 사항이 있을 경우)

그리고 정보자원의 내용에 직접 관련된 메타데이터는 아니지만 데이터의 신규 등록이나 갱신 등과 같은 ORel의 관리용으로 다음의 필드를 만들 수 있다.

등록번호

각 등록 정보자원에 대하여 자동으로 부여되는 번호

등록년월일

정보자원의 등록 연월일

등록자명

정보자원의 등록을 수행한 인물의 이름(웹상의 등록 폼을 사용하여 인터넷상에서 등록된 경우).

등록자 전자우편 주소

정보자원의 등록을 수행한 사람의 전자우편 주소

공개 플래그

정보자원을 공개/비공개로 구분하기 위한 필드

소멸 플래그

정보자원이 소멸되는 경우에 사용되는 필드

ORel에서는 홈페이지 상에서 ORel 관리자 이외의 이용자로부터 정보자원 등록을 받을 수 있다. 등록자명, 등록자 전자우편 필드가 이 경우에 사용되며 정보자원의 등록을 수행하는 이용자의 이름과 전자우편을 입력한다. 공개 플래그는 등록 정보자원을 공개하는 것이 불가함을 지정하기 위한 항목이다.

외부로부터 등록이 이루어지는 경우에도 자동으로 모든 정보자원이 ORel의 등록 정보자원으로서 홈페이지 상에 공개되는 것은 아니며 잘못 기록된 것이 없는지를 확인한 뒤에 운영자 측에서 공개하는 것으로 되어있다. 또한, 등록단계에서는 액세스가 가능한 정보자원이 다음에는 액세스할 수 없는 경우가 있다.

이 경우에 일단 공개 플래그를 사용하여 디렉토리에 표시되지 않도록 한다. 장기간 액세스를 할 수 없게된 정보원은 소멸한것으로 간주하여 활성화된 데이터베이스 항목에서 제외된다. 소멸 플래그는 이러한 용도로 사용된다.

ORel 메타데이터 참조모형

ORel은 앞에서 언급한 대로 등록 정보자원의 메타데이터들을 데이터베이스내에 유지하는 대신 본문의 데이터는 유지하지 않는다. 이를 위해 정보자원과 메타데이터는 항상 별도의 장소에 위치시켜 놓는다. 이 경우에 정보자원과 메타데이터의 관계를 나타내기 위하여 참조되는 양자를 결합시키지 않으면 안된다.

참조에는 데이터 본체로부터 메타데이터를 참조하는 경우와 메타데이터로부터 데이터 본체를 참조하는 경우 등 2가지의 경우가 고려될 수 있다. Warwick Framework에서는 후자의 참조모형으로 외부 참조모형(externally-referenced metadata)과 내부 참조모형(internally-referenced metadata) 등을 소개하고 있다.

내부 참조모형은 메타데이터를 통해 기술된 정보가 데이터의 저자 또는 관리자에 의해 작성되며 데이터 본체의 일부로서 포함된 경우의 메타데이터와 데이터 본체의 관계를 표

현한 것이다. HTML의 메타요소를 이용하는 경우나 도서 판권의 서지정보는 내부참조 메타데이터이다.

반면 외부 참조모형은 메타데이터가 데이터의 저자 또는 관리자와는 무관하게 작성되며 데이터 본체와는 독립적으로 존재하는 경우의 메타데이터와 데이터 본체의 관계를 나타내는 것이다. ORel에서는 메타데이터로부터 문헌정보자원에 대하여 참조가 되고 있기 때문에 메타데이터 참조모형은 외부 참조모형에 해당된다.

현재 웹에서는 원격 메타데이터를 정보자원과 결합시키기 위한 적절한 구조를 사용하고 있지는 않다. 예를 들어 현행 HTML에서 링크정보는 앵커요소〈a〉로서 원래 링크의 정보자원에 들어가 있을 필요가 있다. 앞으로는 XML의 링크기구(XLink 혹은 XPointer)를 이용하여 원격 메타데이터를 일괄적으로 처리하는 방법이 이용 가능하게 될 전망이다.

이 경우 XML의 링크기구를 사용하여 링크 원래의 정보자원, 링크 이전의 정보자원, 링크 정보를 별도로 관리할 수 있기 때문에 ORel과 같은 메타데이터 모형에 기반한 디렉토리서비스는 보다 유연한 메타데이터를 제공할 수 있게 될 것이다.

ORel 관리

앞에서 언급한 대로 ORel에 등록되는 모든 정보자원은 사람에 의해 추출되고 메타데이터와 함께 데이터베이스에 등록된다. ORel 관리자는 정기적으로 웹을 순회하면서 ORel의 취지에 맞는 문헌정보자원을 추출하고 ORel 데이터베이스에 등록하는 수작업을 수행하고 있다.

이러한 작업은 한편으로 비효율적이지만 검색엔진형 디렉토리는 정보검색 노이즈가 높을 뿐만 아니라 문맥으로부터 도출되지 않는 메타데이터를 검색하는 것이 어렵다. ORel의 목표인 양질의 정보에 대한 액세스를 제공하기 위해서는 수작업은 필요한 프로세스이다.

ORel은 등록되는 메타데이터에 관하여 문헌 정보자원의 소재에 관한 정보('유무 여부' 에 관한 정보)와 문헌 정보자원의 특징에 관한 정보("어디에 무엇이 있는가"에 관한 정보)의 2가지를 개념적으로는 구별하고 있다.

ORel에 등록되어 있으면 웹이라는 정보공간에 해당 문헌 정보자원이 존재하는 것이 나타난다. 역으로 말하면 ORel에 등록된 정보자원은 반드시 문헌 정보자원인 것을 보증한다는 의미다.

ORel의 역할중의 하나는 잡다한 콘텐트가 실려있는 웹의 범위를 문헌 정보자원을 포함한 정보공간으로 제한하는 것이다. 이렇게 함으로써 이용자는 논문정보자원의 소재를 탐색하는 프로세스를 생략할 수 있다.

소재정보를 제공하는 것에는 평가 측면을 동시에 가지고 있다. ORel에 등록되어 웹상에 존재하는 것을 나타내는 문헌 정보자원을 선택하고 ORel에 등록함으로써 정보자원을 평가하는 소기의 목적에 합치되는지를 판단하는 것이다. 정보자원을 평가하지 않고 정보자원을 등록하면 궁극적으로는 웹상의 전 문서가 ORel에 등록되어 결과적으로 정보공간을 압축하여 한번에 양질의 정보에의 억세스를 제공한다는 당초의 목적을 달성하지 못하게될 가능성이 있다. 정보자원의 평가는 디렉토리 사이트에 있어서 매우 큰 문제이다.

이 때문에 ORel에서는 정보원의 평가를 독자적으로 수행하지 않고 외부의 평가를 활용하여 학회지나 전문서적으로 출판되는 문헌을 대상으로 한다. 이렇게 함으로써 평가에 관한 부분을 외부화, 분산화하여 ORel의 데이터 신뢰성을 높이고 대신 웹 정보자원의 메타데이터제공에 전력을 기울일 수 있게 된다.

ORel에 등록된 메타데이터에는 콘텐트나 컨텍스트로부터 도출될 수 있는 내재적 메타데이터와 콘텐트나 컨텍스트로부터 도출될 수 없는 외재적 메타데이터가 있다.

내재적 메타데이터에는 제목, 저자명, 발표년월일 등이 있다. 외재적 메타데이터에는 해당 정보자원에 관한평가, 분류 등이 있다.

내재적 메타데이터는 메타데이터의 등록작업 수행과정상에서 콘텐트의 본체에 표시되는 명시적인 메타데이터와 콘텐트의 내부에는 직접 포함되지 않지만 해당 정보자원이 작성되어 공개된 컨텍스트에 따라 함의된 암묵적 메타데이터로 구분된다.

예를 들어, 어느 URL에는 정보자원의 제목이 기재되고 저자명이 기재되어있지 않다고 한다. 정보자원의 본체에 저자명이 기재되어있지 않더라도 관련 페이지(대부분의 경우

에는 색인 페이지)를 참조함으로써 저자명을 알 수 있다. 이 경우에 저자명은 함의된 내재적 메타데이터이다. 현실적으로 메타데이터를 이렇게 명확하게 구분짓는 것이 어려울 수도 있지만, 이론적으로는 이렇게 구분된다. 이러한 관계를 〈그림〉에 표시한다.

〈그림〉 메타데이터 분류

웹상의 문헌 정보자원에는 메타데이터가 명시적으로 기록되어 있는 경우가 많기 때문에 ORel 구축에서 중심되는 작업은 〈그림〉의 암묵적 내재 메타데이터를 도출하는 작업이다. 문헌 정보자원의 페이지 상에 필요한 메타데이터 항목이 기재되어 있는 경우도 많지만 예를 들어 색인 페이지에 저자명, 논문명 등의 메타데이터가 기재되어 있어도 논문 정보자원 본체의 페이지에는 정보가 표시되지 않는 경우가 있다.

이러한 경우에 기재되어 있지 않은 저자명, 논문명 등의 정보는 명시적으로는 기재되어 있지 않지만 문헌 정보자원에 내재되어 컨텍스트로부터 판별하는 것이 가능하다. ORel을 작성하는 경우에는 각각의 정보자원 마다 암묵적으로 내재하는 메타데이터를 추출하고 있다.

색인을 위한 페이지와 콘텐트를 위한 페이지는 식별자(URL)의 관점에서는 서로 독립적인 엔티티이기 때문에 의미있는 모양으로 결합되어 있지 않다. 왜냐하면 우연히 근접한 URL을 가지고 있을 뿐이다.

콘텐트 페이지에 쓰여있지 않은 저자명을 색인 페이지에서 추출하는 작업도 빈번히 발생한다. 이것도 정보자원의 컨텍스트로부터 도출되는 암묵적인 메타데이터라고 말할 수 있다. 자동색인에서 이러한 두 페이지의 관계까지 추론하여 도출할 수 없기 때문에 수작업이 유효하게 기능하는 측면이 있다.

동시에 이것은 현재의 HTML/HTTP의 웹 환경이 갖는 제약이기도 하다. 이러한 제약에 따라 현재는 여러개의 URL로 분할되는 객체의 상호관련성을 외부적으로 표시하는 것이 불가능하다. XML의 링크기구를 이용하면 메타데이터와 본체의 경우와 마찬가지로 독립된 URL을 갖는 두 개의 앤티티간의 링크정보만을 분리하여 유지하는 것이 가능하기 때문에 색인 페이지와 콘텐트 페이지 간의 관계를 기술하는 것도 가능하다.

또한 전문검색형의 검색서비스에서는 정보자원중에 발생하는 어구를 바탕으로 하여 '논문', '전자잡지', '통계자료'와 같은 정보자원의 종별에 따른 검색을 수행하는 것도 불가능하다. 이것은 어느 정보자원이 논문인지 아닌지 본문에서 추론할 수 없는 외적 메타데이터에 소속되어 있기 때문이다.

결론

ORel은 웹사이트의 Yahoo와 마찬가지로 사람이 웹상에 있는 학술 정보자원을 검색한 후, 메타데이터를 선택하여 데이터베이스에 등록하고 있다. ORel 처럼 수동 색인형 서비스는 데이터의 등록에 많은 인력을 필요로 하기 때문에 갱신이 빈번하게 이루어지는 웹 전체를 색인 대상으로 하는 것은 올바르지 않다. 웹 전체를 대상으로 포괄적인 검색을 하는 것은 자동검색엔진을 사용하는 전문검색형 서비스에서는 유효하다.

그러나, ORel이 대상으로 하는 학술목적의 문헌 정보자원의 경우에는 학술문헌이라는 정보자원의 성격상 일반적인 웹 정보자원과 다르며 갱신의 빈도는 비교적 낮다. ORel을 운영하는 동안 URL이 이동하는 정보자원은 약간 나타나기는 하지만 약간의 변경을 제외하고는 내용이 대폭 갱신되는 문헌 정보자원은 발견되지 않았다. 이와같은 정보자원의 성격 때문에 ORel에서는 사람이 직접 데이터베이스를 관리하고 있으나 웹의 빈번한 갱신에 대응할 수 없다는 문제점을 가지고 있다.

이상 웹상의 문헌 정보자원 검색서비스인 ORel에 대하여 소개하였다. ORel은 범용성이 있는 형태로 메타데이터를 기술하고는 있지만 현시점에는 다른 유사한 서비스와의 상호운용은 실현되고 있지않다.

이것은 앞으로 해결해야 할 과제이다. 이 경우에 ORel은 다른 데이터베이스를 집약하는 서비스로서 뿐만 아니라 미래에는 이들 서비스를 통합하는 메타서비스의 하나의 구성요소로서 기능을 하게 될 것이다.