제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

학술논문 메타데이터 변환 도구의 설계

이민호, 이원구, 윤화묵, 성원경
한국과학기술정보연구원
e-mail:cokeman@kisti.re.kr

A Design of Metadata Conversion Tool for Research Paper

Min-Ho Lee, Won-Goo Lee, Hwa-Mook Yoon, Won-Kyung Sung
Korea Institute of Science and Technology Information

요 약

대량의 데이터를 분석하여 보다 차원 높은 정보서비스를 제공하기 위해서는 다양한 데이터의 통합 관리가 필수적이다. 특히 과학기술 분야에서는 논문 메타데이터를 분석하여 연구동향 파악, 선도 연구자 파악 등을 하기 위한 연구가 진행 중이다. 논문 메타 데이터의 통합 관리를 위해서는 메타데이터 스키마의 매핑과 데이터 변환이 필요한데, 본 논문에서는 논문 메타데이터 변환에서의 문제를 분석하여 보고, 해결하기 위한 방법을 제시하였다. 또한 다양한 구문을 지원하면서 스키마에 유연하여 시스템 수정이 필요없는 도구를 설계하였다.

1. 서론

최근 정보서비스의 경향은 기존의 단편적인 정보검색 및 제공 서비스를 벗어나 다양한 정보원으로부터 대량의 정보를 수집 및 분석하여 보다 유용한 정보를 제공하려고 하고 있다. 과학기술 분야에서도 과학기술 문헌으로부터 글로벌 연구동향의 파악, 이머징 시그널 탐지, 선도연구자 파악 등을 하기위한 많은 연구가 수행 중이다. 이러한 분석을 위해 수집되는 과학기술 문헌정보는 각 정보원 자신들의 목적에 맞도록 생산되었기 때문에 구성과 표현이 서로 상이하며 유용한 정보의 분석 및 추출을 용이하게 하기 위해서는 같은 구조와 표현형식으로 통합하여 관리하여야 한다.

본 논문에서는 통합 관리를 위해 다양한 형태의 학술논문 메타데이터를 통합 메타데이터로 변환 시 고려하여야 할 문제를 살펴보고 이에 특화된 변환 도구를 설계한다. 2장에서는 메타데이터 변환과 관련된 연구에 대해서 알아보고, 3장에서는 학술논문 메타데이터 변환시 고려할 문제를 분류해 본다. 4장에서는 학술논문 메타데이터 변환을 쉽게 할 수 있는 도구를 설계하고, 끝으로 5장에서 결론을 맺는다.

2. 관련연구

통합 관리 및 분석 추출의 대상이 되는 논문의 메타데이터는 사람에 따라 정의가 조금씩 다르나 Priscilla Caplan[1]은 '어떤 정보 자원에 관한 구조화된 정보를 의미하는 것'으로 정의하고 메타데이터 스키마를 구문, 의미구조, 내용 규칙 세 측면으로 분류하였다. 구문은 메타데이터를 이루는 요소들을 어떻게 기계 가독 형식으로 인코딩해야하는 가를 말한다. 의미구조는 메타데이터 요소 자체의 의미를 말하는 것으로 일반적으로 이름과 정의로서 의미를 표시한다.

서로 다른 메타데이터 스키마간의 의미구조 호환성을 위해서는 일반적으로 크로스워크[2], 자동 혹은 수동 스키마 매핑[3], 메타데이터 레지스트리[4]를 통하여 해결한다. 내용규칙은 메타데이터 요소의 값이 어떻게 선정되고 표현되는지를 명시한다.

메타데이터 변환을 위한 연구로는 XSLT를 이용하여 XML 형식의 데이터를 변환하는 방법[5]과 스키마 매칭과 내용 규칙 변환을 위한 함수들을 설정하는 인터페이스를 제공하여 변환 프로그램 모듈을 자동으로 생성하는 도구[6] 등이 있었다.

상기 도구들은 컨텐츠 독립적인 범용 도구이기는 하나 XML 문서만을 변환 대상으로 하는 구문 의존적이거나 다양한 변환 함수를 지원하기 위하여 사용법이 복잡하였다. 또한 자동 생성된 프로그램 모듈을 컴파일하여 변환 응용 프로그램을 작성하여야 하기 때문에 새로운 스키마를 가진 메타데이터를 추가하려면 재컴파일하여야 하는 번거로움이 있었다.

3. 학술 논문 메타데이터 변환 문제

논문의 개략적인 내용 파악과 검색을 위해 주로 사용되는 메타데이터는 출판사별로 상이한데, 출판사 나름대로의 용도에 따라 자체적으로 정의하여 사용하기도 하나 주로

Dublin Core, MARC, ISO12083 등의 메타데이터를 확장하거나 변형하여 사용하는 편이다.

기계가독 형식으로 표현하기위한 메타데이터 구문은 SGML, XML이 주로 많지만 자체 정의한 Tag를 이용한 일반 텍스트(이하 Tagged Text)와 HTML도 존재한다. 따라서 변환 도구는 다양한 구문의 메타데이터 파서를 갖추어야 한다.

```
@--
_ti Special Issue: Atmospheric Nitrous &cOxide
_au Khalil, M.A.K.
_ab Importance of this paper: we report ...
_vi 2
_is 3-4
@--
```

(그림 1) 서지 메타데이터의 예

그림 1은 과학기술 분야의 거대 출판사인 한 출판사에서 제공하는 서지 메타데이터의 일부로서, Tagged Text 형태이며 '@--'는 한 파일에 다수의 메타데이터가 있을 경우 각 메타데이터를 구분하기 위한 문서 구분자이다. '_'로 시작하는 요소명(태그)을 가지고 있으며, 공백이후 해당 태그의 값이 들어가 있다. 이 예에서 '_ti'는 논문의 제목을 나타내는 태그이며, '_au', '_ab', '_vi', '_is'는 각각 저자명, 요약, 볼륨번호, 이슈번호를 나타낸다.

메타데이터의 의미구조와 내용규칙 측면에서의 변환 문제를 설명하기 위하여 변환되어야 하는 통합 메타데이터(이하 목적데이터)의 스키마가 XML로 기술되며, 위 데이터의 각 태그를 그림 2로 표현한다고 가정하자.

```
<article-meta/article-title>
   Special Issue: Atmospheric Nitrous Ōxide
</article-meta/article-title>
<article-meta/surname>
   M.A.K.
</article-meta/surname>
<article-meta/given-name>
   Khalil
</article-meta/given-name>
<article-meta/volume>
   Vol. 2, No. 3-4
</article-meta/volume>
```

(그림 2) 통합 메타데이터의 예

목적 메타데이터와 원시 메타데이터의 차이점은 저자명이 원시 메타데이터에서는 성과 이름이 하나의 요소로 기술되어 있는데 반해 목적 메타데이터에서는 두 개의 요소로 분리되어 있으며, 반대로 볼륨번호, 이슈번호 두 개의 요소는 <article-meta/volume> 하나의 요소로 합쳐져 있다. 요소명 및 요소의 구조 뿐만 아니라 데이터 값 또한 변환되어야 한다. 원시데이터는 문자표기를 위하여 KSC-5601 코드를 사용하였기 때문에 '_ti'의 경우 '&c'를 사용하여 'Ō'를 표현하였지만, 목적데이터는 UTF-8 코드를 사용하여 'Ō'를 직접 표현하였다. 또한 '_au' 요소의 값인 Khalil, M.A.K. 는 컴마(,)를 구분자로 하여 두 개의 요소 값으로 분리하고 컴마는 버리며, '_vi', '_is' 요소의 값은 'Vol'., 컴마(,), 'No.'를 사용하여 하나의 값으로 합쳐야 한다.

위의 예와 같이 요소명, 요소의 구조, 문자 코드, 문자열의 추가 혹은 삭제, 다른 형태로의 문자 변환 등 서지메타데이터의 변환에는 여러 가지 사항을 고려하여야만 한다. 이러한 논문 메타데이터 변환의 문제를 메타데이터 요소의 의미, 구조와 연관된 문제들은 스키마의 이질성으로 내용 규칙과 관련된 문제들은 데이터의 이질성으로 다음과 같이 분류하였다.

1) 스키마 이질성 관련 변환

1-1) 이름 변환 : 같은 이름으로 사용된 항목이 서로 다른 개념을 표현하거나 서로 다른 개념을 표현하는데 같은 이름으로 사용되는 경우이다.
예) _date(최종수정일) -> <article-date>(논문 출판일)

1-2) 구조 변환 : 원시 스키마(변환 전의 스키마)와 목적 스키마(변환 후의 스키마)의 구조가 다를 때 발생한다.
예) _au -> <article-meta/surname>,
<article-meta/given-name>

2) 데이터 이질성 관련 변환

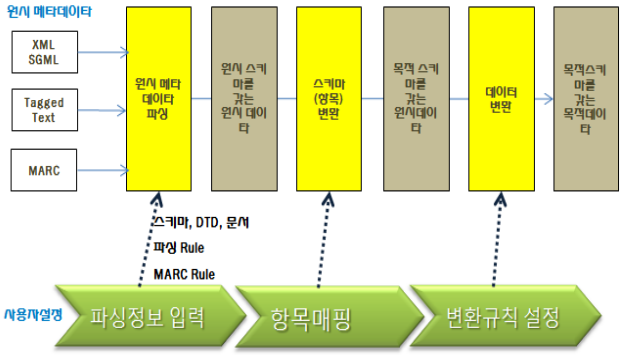
2-1) 코드 변환 : 다른 국가코드, 날짜표기 코드 등 관리 혹은 데이터 전송의 편리함 때문에 코드로 관리되는 것들의 변환이다.
예) 1월 -> Jan , China (IEEE 국가코드의 중국) -> CHN (ISO3166 국가코드의 중국)

2-2) 패턴 변환 : 값이 일정한 문자열 패턴을 가지고 있는 경우이다. 주로 문서번호, 저자명의 성과 이름, 발행 권호, 날짜, 페이지 등의 요소들이다.
예) 13th edition, V.1 -> Vol. 1 no. 13

2-3) 기타 변환 : 코드나 패턴으로 정의되지 않는 불규칙한 내용변환이 일어나는 경우이다. 예를 들면 논문 기사 페이지 범위를 페이지 수로 변환해야 하는 경우나 요소 내용의 문자열 중 의미 없는 부분을 버려야 하는 경우이다.
예) i-v (페이지 범위) -> 5 (페이지 수)

4. 변환 도구의 설계

설계한 변환 도구에서는 메타데이터 변환 과정을 스키마 이질성을 해결하는 스키마 변환과정과 데이터 이질성을 해결하는 데이터 변환과정으로 나누어 문제를 단순화하고 사용자가 스키마 이질성을 해결하는 스키마 매핑과 데이터 값 변환 규칙을 정할 수 있도록 하였다.



(그림 3) 메타데이터 변환 과정

그림 3은 메타 데이터 변환 과정을 도식으로 표현한 것으로 순서대로 상세히 설명하면 다음과 같다.

1) XML, SGML, Tagged Text, MARC 등 다양한 구문의 원시 메타 데이터가 입수되면 사용자는 해당 데이터의 파싱정보를 시스템에 입력한다. 파싱정보는 원시메타데이터의 형태에 따라 XML Schema, XML DTD, SGML Schema,MARC Rule 등이며 Tagged Text인 경우에는 문서의 구조 정보를 직접 입력할 수도 있다. 사용자가 입력한 파싱정보를 바탕으로 원시 메타데이터를 파싱하여 원시 스키마를 갖는 원시 메타 데이터가 생성된다.

2) 스키마 변환단계이다. 사용자는 원시 스키마의 각 요소(항목)들을 목적 스키마의 각 항목들과 매핑한다. 시스템은 매핑된 정보를 이용하여 스키마 변환을 수행한다. 이때 각 항목의 데이터는 변하지 않고 매핑 유형에 따라 1:1로 매핑되는 경우는 그대로 복사되고 하나의 항목이 여러 개로 분리되는 1:N 매핑이거나 여러 개의 항목이 하나로 합쳐지는 N:1 매핑인 경우에는 구분자를 가지고 여러 항목의 데이터를 합치거나 한 항목의 데이터를 분리한다. 이 과정을 통하여 스키마 이질성이 해결되며, 시스템은 목적 스키마를 갖는 원시데이터 임시 데이터베이스를 생성한다.

3) 데이터 변환 단계로서 사용자는 목적 스키마 항목의 의미와 형식에 맞도록 데이터 변환을 위한 규칙을 설정한다. 규칙 설정은 3장에서 메타데이터 변환 시 데이터 이질성의 문제로 정리한 각 변환에 따라 여러 가지 방법으로 설정하여 변환한다. 각 변환 방법은 다음과 같다.

3-1) 코드변환

코드변환은 미리 코드 테이블을 모두 등록하여 두고 변환할 코드와 변환대상이 되는 코드를 사용자가 지정하게 하여 편리하게 변환할 수 있다.

3-2) 패턴변환

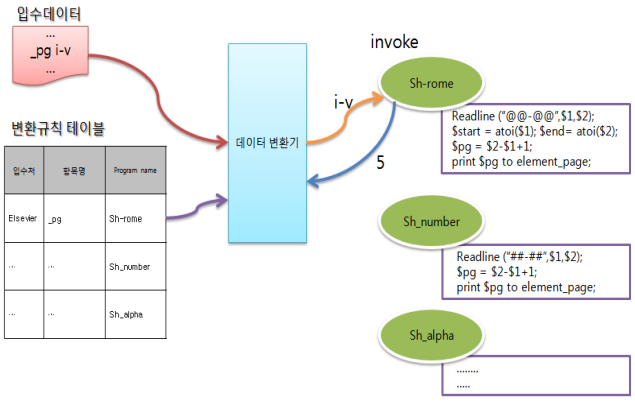
패턴변환의 형태는 문자열 패턴을 사용자가 입력토록 하여 변환한다. 예를들면, 볼륨번호가 1998이고 이슈번호가 13인 책의 권호 항목이 v.1998 no.13 이라고 표시되어 입력된다고 가정하자. 이것은 v.@@@@ no.## 인 패턴으로 시스템은 @@@@와 ##만을 추출해 낼 수 있다. 사용자가 출력 패턴을 Vol.@@@@ no.## 이라고 설정하면 시스템은 추출한 @@@@와 ##에 Vol.과 no.만을 문자열 추가를 하여 변환할 수 있다.

3-3) 기타변환

기타변환은 입력 파라미터 1개와 출력 파라미터 1개를 가지는 스크립트 언어로 표현된 API를 시스템에 등록하게 하여 수정없이 변환할 수 있다. 전형적인 예로, 논문집에서 특정 논문이 들어있는 페이지를 나타내기 위하여 원시 항목에서는 시작 페이지와 끝 페이지를 로마자를 이용하여 i-v와 같은 형태로 표시하고 목적 항목에서는 단지 페이지의 수만을 표시한다고 하자. 이와 같은 변환은 코드 테이블을 이용하거나 패턴등록을 통하여 변환하기 어렵다. 대신 변환과정을 스크립트 언어를 사용하여 그림 3의 예와 같이 기술할 수 있다.

```
Read ("@@-@@",$1,$2);
$start = atoi($1); $end = atoi($2);
$pg = $end - $start + 1;
print $pg to page element of destination database;
```

등록한 스크립트 언어는 그림 4와 같은 방식으로 호출되어 실행된다.



(그림 4) 기타변환 규칙화면에서 정의된 스크립트 실행

앞에서 기술한 3가지 형태의 데이터 변환을 통하여 새로운 형태의 논문 메타데이터가 입수되더라도 시스템의 수정 없이 사용자 설정만으로 메타데이터의 변환이 가능하다.

5. 결 론

본 논문에서는 과학기술정보의 통합관리를 위하여 다양한 형태의 학술논문 메타데이터를 통합 메타데이터로 변환할 때의 문제를 분류하여보고 학술논문에 특화된 변환 도구를 설계하였다. 설계한 도구는 다양한 형식의 구문을 지원하며, 3가지 형태의 데이터 변환 방법을 통하여 새로운 형태의 논문 메타데이터도 시스템의 수정없이 사용자 설정만으로 변환이 가능하다.

본 도구에서는 스키마 변환을 위한 매핑을 사용자가 수동으로 실행하나 다수의 요소를 가진 스키마 매핑을 쉽게 하기 위하여 자동 스키마 매핑 추천 서비스에 대한 연구를 향후 진행할 계획이다.

참고문헌

[1] Priscilla Caplan, 오동근 역, "메타데이터의 이해", 태일사, 2004.

[2] Margaret St. Pierre, "Issues in Crosswalking Content Metadata Standards", NISO White Paper, 1998.

[3] Erhard Rahm, Phillip A. Bernstein, "A survey of approached to automatic schema matching", The VLDB Journal, Vol.10, pp.334-350, 2001.

[4] Metadata Registry, http://metadata-stds.org/11179/

[5] 하태진, "RCP를 이용한 XSLT 기반의 스키마 매퍼 구현", 석사학위 논문, 서울과학기술대학교, 2009.

[6] Altova Mapforce, http://www.altova.com/mapforce.html