의미 호환을 위한 메타데이터 매핑 연구

A Study on Metadata Mapping for Semantic Interoperability

고 영 만 (Young Man Ko)*
서 태 설 (Tae-Sul Seo)**
임 태 훈 (Tae-Hoon Lim)***

초 록

본 연구에서는 다양한 메타데이터간의 의미적 호환성을 유지하거나 개선하기 위한 기존의 방법론을 분석하고 크로스워크를 이용한 메타데이터간의 의미 호환 가능성과 한계에 대해서 검토한 후 메타데이터간의 의미 호환을 극대화하기 위한 의미적 메타데이터 매핑 프로세스를 제시하였다. 이 프로세스는 대상 메타데이터 스킴 확인, 공통 데이터요소개념(DEC) 발견, 데이터요소개념에 따른 속성 그룹화, 매핑 테이블 작성 등의 네 단계로 구성된다. 국내에서 개발된 단체표준 수준의 두 인력정보 메타데이터를 대상으로 본 연구에서 제안된 프로세스를 적용하여 매핑 테이블 작성 과정을 보였다.

ABSTRACT

This paper contains an analysis of the methods that have been used to achieve or improve interoperability among metadata and discuss the possibilities and limits of semantic interoperability among metadata using crosswalk. After that a semantic metadata mapping process which is able to maximize the interoperability among metadata is suggested. The methodology consists of four steps such as identifying metadata schema, finding common data element concepts(DECs), grouping attributes by the DECs, and mapping into a table. An experimental application of the process was performed onto two human resource information metadata standards developed in Korea.

키워드: 메타데이터, 메타데이터 크로스워크, 의미 호환, 매핑 테이블, 인력 정보
Metadata, Metadata Crosswalk, Metadata Registry, Semantic Interoperability, Mapping Table, Human Resource Information,

---

논문제출일자 : 2007년 11월 15일
* 성균관대학교 문헌정보학과 교수(ymko@skku.ac.kr)
** 한국과학기술정보연구원 지식전략팀 책임연구원(tsseo@kisti.re.kr)
*** 한국데이터베이스진흥센터 지식표준팀 선임연구원(taehoon@dpc.or.kr)

1. 서 론

이 세상에는 이미 다양한 메타데이터 셋이 만들어져서 통용되고 있으며 심지어 동일 기관 내의 동일 정보 자원에 대해서도 다른 메타데이터 스킴을 사용하는 경우가 있다. 이는 데이터베이스의 개발자에 따라 동일한 데이터 요소(data element)에 대해서 다른 이름을 부여했기 때문이며, 이로 인해 대부분의 경우 데이터의 의미적 불일치(semantic heterogeneities)가 발생한다. 따라서 메타데이터 차원에서의 의미적 호환성(semantic interoperability)은 동일하거나 유사한 도메인에서의 정보 교환과 통합 요구가 발생할 경우 우선적으로 해결해야 할 과제로 간주되고 있다. 이와 관련하여 서태설 등(2007)은 의미적 불일치의 유형을 분석하고 각 유형별 의미 일치 방안을 찾아내는 작업이 매우 중요한 것임을 강조한 바 있다.

메타데이터의 상호운영을 확보하기 위한 대표적인 방법으로는 크게 메타데이터 크로스워크를 만들어 사용하는 방법, 자원의 속성을 감안하여 다양한 메타데이터 형식과 기술 구조를 인정하고 상호 매핑을 통해 해결하는 방법, 그리고 메타데이터 레지스트리(metadata registry : MDR)에 의한 해결 방법의 세 가지를 들 수 있다. 이 가운데 메타데이터 레지스트리에 의한 방법과 메타데이터 크로스워크에 의한 방법은 의미적 일치를 위한 해결에 있어서 매우 대조적인 방식으로 접근한다. 메타데이터 레지스트리에 의한 방법은 중앙의 메타데이터 레지스트리를 구축하여 통제하는 방법으로 특정 도메인에서 메타데이터 표준을 서로 공유함으로써 사전 대비를 추구하는 접근 방식을 보이고 있으며, 메타데이터 크로스워크에 의한 방법은 이미 통용되고 있는 상이한 메타데이터 간의 의미 호환을 확보함으로서 기존의 데이터를 교환하거나 통합하는 사후 조치적 접근 방법을 보이고 있다 (고영만 2005).

데이터의 의미적 상호운용성 문제를 심각하게 인식하고 본격적으로 다룬 해외의 초기 연구로는 환경 데이터의 의미적 상호운용성 확보를 위하여 분산 에이전트 아키텍처를 개발한 “InfoSleuth 프로젝트”를 들 수 있다 (Fowler et al. 1999). “InfoSleuth 프로젝트”에서는 개발자가 대상 분야의 개념과 개념 간의 관계를 낮은 수준의 데이터베이스 스키마로 이전할 수 있는 용어로 표현하게 함으로써 사용자 간의 의미적 교환을 가능하게 하는 방식을 제시하였다. 국내에서 데이터 간 의미적 상호운영성 문제를 메타데이트 레지스트리 측면에서 해결하고자 한 연구는 심경(2003)에 의해 처음 시도되었다. 그는 이종 메타데이터의 통합 틀로서 "RDF(resource description framework, 이하 RDF)"와 메타데이 레지스트리를 비교 조사함으로써 메타데이터 통합 방법으로서의 RDF와 메타데이터 레지스트리의 가능성을 분석하였다. 실제 적용하는 것과 관련된 것으로는 고영만과 서태설(2005)이 서지정보를 대상으로 데이터 요소의 명명규칙 방법론을 제시한 연구가 있다. 이 연구에서는 정보자원의 메타데이

터를 작성할 때 메타데이터 간 의미의 일관성 유지에 적합한 메타데이터의 명명 방법론을 제안하고, 이를 실제 분야에 적용할 수 있도록 메타데이터 명명 규칙의 실험적 모형을 제시하였다. 또한 서태설과 Pham(2007)은 "ISO/IEC 11179 – MDR" 표준에 근거하여 기존의 데이터 모델과 메타데이터 레지스트리를 혼합한 의미적 데이터 모델링 프로세스를 개발하였다.

메타데이터의 의미적 상호운용성 문제를 크로스워크(crosswalk) 측면에서 다룬 연구로는 유에스마크(USMARC), 엘오엠(LOM), 더블린코어(Dublin Core)의 세 개 메타데이터에 대한 크로스워크 작성을 "XML"의 네임스페이스 수준에서 시도한 라이트 등(Light et al. 2003)의 연구를 들 수 있다. 또한 더블린코어와 마크를 대상으로 갓비 등(Godby et al. 2004)이 메타데이터 크로스워크 저장소 모델을 제시하여 "METS(Metadata Encoding and Transmission Standard)"를 구현한 연구가 있다. 이 외 해외에서는 미의회도서관, "OCLC", "UKOLN", "Getty연구소", "DLESE 프로젝트" 등에서 메타데이터 크로스워크에 대한 연구와 개발을 하고 있으며(http://www.slis.kent.edu/~mzeng/metadata/crosswalks.htm 참조), 국내에서는 한국데이터베이스진흥센터의 메타데이터 크로스워크 포럼을 중심으로 관련 연구가 활성화되고 있다.

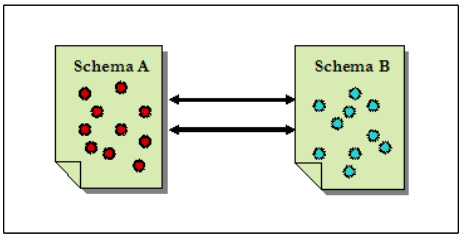
그렇지만 크로스워크와 관련하여 시도된 대부분의 연구는 단순한 매핑 측면에서 메타데이터의 상호 교환 가능성을 제시하는 수준에 머물러 있으며, 메타데이터 크로스워크 방법론을 기반으로 의미적 호환성을 분석하고 방법론을 체계화하려는 심도 있는 노력은 아직까지 찾아보기 힘들다. 따라서 본 연구에서는 크로스워크를 이용한 메타데이터 간의 의미 호환 가능성과 한계의 검토 및 "ISO/IEC 11179-3"에서 제시된 레지스트리 메타모델에 근거하여 메타데이터 간의 의미 호환을 극대화할 수 있는 의미적 메타데이터 매핑" 프로세스를 제안하고자 하며, 최소한 국내의 단체표준 수준으로 개발된 두 인력정보 메타데이터 표준을 대상으로 하여 본 연구에서 제안하는 프로세스의 적용사례를 제시하고자 한다.

2. 메타데이터 크로스워크

2.1 메타데이터 크로스워크 방법론

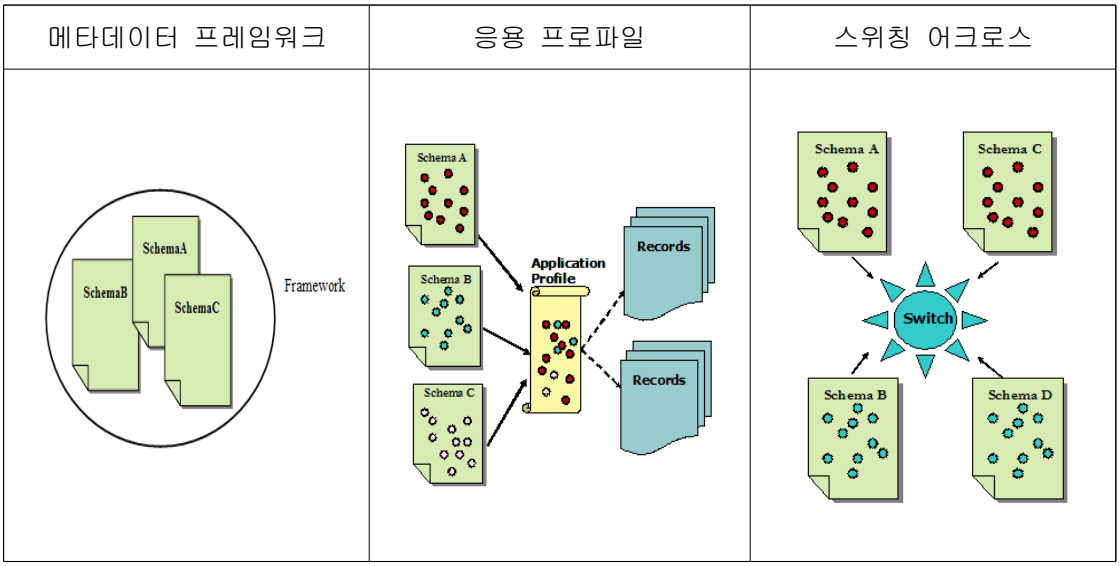
메타데이터 크로스워크는 여러 메타데이터 간에 메타데이터 요소의 의미와 구조를 매핑하는 것으로서, 오늘날 메타데이터 간의 상호운용을 위해서 가장 많이 사용되는 방법이다. 크로스워크 방법에서는 하나의 "source"에 해당하는 독립적인 메타데이터 셋을 "target"에 해당하는 상대 메타데이터 셋에 데이터 요소 수준에서 매핑을 하게 되며, 그 결과로 하나의 데이터 요소 매핑 테이블 또는 차트를 생성시킨다 (그림 1 참조). 이러

한 메타데이터 크로스워크는 잘 알려진 메타데이터 표준 간에 만들어질 수 있으며, 동일 표준의 다른 버전 간에도 만들어질 수 있다.



<그림 1> 메타데이터 크로스워크 개념(Chan et al. 2006)

크로스워크의 생성 방식에는 절대적 크로스워킹과 상대적 크로스워킹 두 가지가 있다. 절대적 크로스워킹은 상대 요소와의 완전 매핑을 하는 방식이며, 상대적 크로스워킹은 모든 대상 데이터 요소를 최소한 하나 이상의 상대 요소에 매핑을 하는 방식으로 완전 매핑을 목표로 하지는 않는다. 한편 찬 등(Chan et al. 2006)은 넓은 의미에서 메타데이터 크로스워크에 해당하는 메타데이터 간 의미호환 방법을 메타데이터 프레임워크(metadata framework), 응용 프로파일(application profile), 스위칭 어크로스(switching-across)의 세 유형으로 나누어 분석하였다. 메타데이터 프레임워크는 메타데이터의 프레임을 만들고 그 프레임에 다양한 메타데이터 요소들을 체계화 하는 방식이며, 응용 프로파일은 여러 개의 메타데이터 셋으로부터 특정 응용 분야에 해당하는 데이터 요소만을 하나의 응용 프로파일로 만들어서 사용하는 방식이다. 그리고 스위칭 어크로스는 하나의 메타데이터 셋이 기준이 되어서 여러 개의 메타데이터 셋과 매핑 함으로써 여러 개의 메타데이터 셋이 서로 간에 매핑하지 않아도 되게 하는 방법이며, 데이터 요소의 의미 분석에 온톨로지 개념을 적용할 경우 매핑의 수준이 매우 높아질 수 있는 방법론이다 (그림 2 참조).



<그림 2> 메타데이터 크로스워크의 방법론(Chan et al. 2006)

2.2 크로스워크의 한계와 가능성

크로스워크는 메타데이터간의 의미적 호환을 확보하는데 있어서 비교적 용이하고 효과적인 방법이 될 수 있으나 완전한 1대1 매칭이 안되는 경우가 자주 발생하는 문제점을 가지고 있다. 하나의 요소가 상대방의 요소와 중첩되는 경우, 두 개로 분리되는 경우 또는 아예 매칭이 되지 않는 경우가 전형적으로 나타나는 문제점이며, 매핑 과정에서 이러한 문제가 발생하게 되면 매핑의 질이 저하되는 것을 피할 수 없게 된다. 메타데이터 매핑은 원칙적으로 1대1 완전 일치가 될 경우에만 자동적으로 의미적 상호 호환을 할 수 있으므로 지금까지 많은 크로스워크가 개발되었으나 실제에서는 큰 도움이 되지 못하는 한계를 가지고 있다. 이와 관련하여 세인트 피에르 등(St. Pierre et al. 1998)은 미국의 국가정보표준기구(National Information Standards Organization : NISO)에서 발간된 보고서에서 메타데이터 간의 매핑을 자동으로 수행하기 위해서는 공통 용어의 공유, 조직화 방법 및 개발 절차의 일치가 중요함을 지적한 바 있다.

그렇지만 크로스워크는 매핑에 따른 전형적 문제점이 해결될 경우 관련 도메인 내의 메타데이터 간 의미적 관계성을 분석할 수 있는 좋은 도구이며, 메타데이터 프레임워크, 스위칭 어크로스의 개념은 데이터 요소 간의 개념 관계를 의미적으로 분석할 경우 매우 유용하게 사용될 수 있는 방법이다. 특정 도메인 내에서 존재하는 여러 메타데이터 셋에 대한 개념 분석을 통해서 프레임화 하고 특정 기준 메타데이터 셋을 중심으로 매핑할 경우 각 데이터 셋들의 의미적 유사도를 확인함으로써 메타데이터의 의미적 상호 호환 작업의 근간이 될 수 있다. 따라서 본 연구에서는 "ISO/IEC 11179-MDR" 표준에서

제시하는 메타데이터 구성 원리를 바탕으로 스위칭 어크로스 개념의 이른바 “의미적 메타데이터 매핑(semantic metadata mapping : SMM)” 프로세스를 개발하여, 이 프로세스에 따라 적용 사례를 제시하고자 한다.

3. 의미적 메타데이터 매핑

3.1 의미적 메타데이터 매핑 프로세스

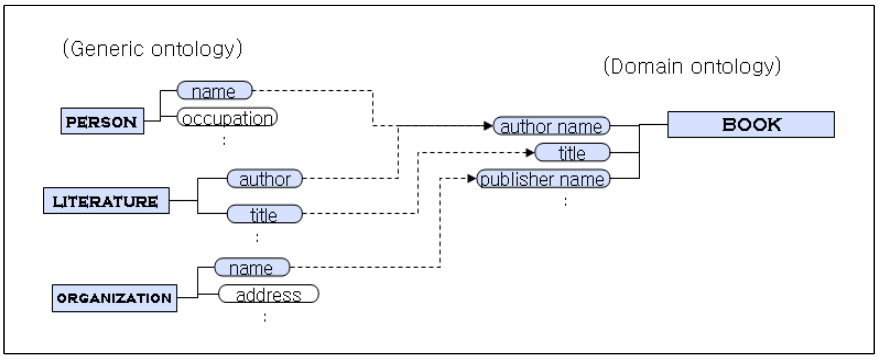
일반적인 메타데이터 크로스워크가 주로 1:1 매핑에 사용되는데 비해 스위칭 어크로스는 여러 개의 메타데이터 셋을 매핑하는데 적합하다. 특히 스위칭 어크로스 방식은 특정 메타데이터 셋을 기준으로 삼아 매핑을 지원하기 때문에 매핑하고자 하는 메타데이터 셋 간의 개념 분석이 필요하게 된다. 따라서 본 연구에서 제시하는 의미적 메타데이터 매핑 프로세스는 스위칭 어크로스 방법론을 토대로 의미적 분석 과정에 ISO/IEC 11179 표준에서 제시하고 있는 메타데이터 모델을 적용하여 그 과정을 특성별로 단계화한 것이다. 이 때 의미적 분석 과정에 메타모델에서 제시하는 데이터 요소의 개념과 개념 간의 관계, 데이터요소개념과 속성(property) 간의 관계, 데이터요소와 속성의 재현(representation) 등을 고려하여 매핑 과정을 단계화 하였다. 이렇게 해서 얻어진 전체적인 프로세스는 대상 메타데이터 스킴 확인, 공통 데이터요소개념(data element concept : DEC) 확인, 데이터요소개념에 따른 속성의 그룹화, 매핑 테이블 작성의 네 단계로 구성된다.

대상 메타데이터 스킴의 확인 단계에서는 의미적으로 상호 호환하고자 하는 대상 메타데이터 셋의 범위를 정한다. 범위의 확정을 위해서는 해당 도메인에서 가용한 모든 메타데이터를 조사하여야 하며, 일정한 양식을 정하여 메타데이터 간 비교가 가능하도록 조사하는 것이 중요하다. 이때 고려하여야 할 사항은 해당 메타데이터가 어떤 객체를 다루고 있는가, 사용되는 응용 영역은 어디인가, 데이터 요소의 수는 몇 개인가, 샘플 데이터는 무엇인가 등이다. 더블린코어, 마크, “GILS”의 세 메타데이터를 대상으로 메타데이터 스킴을 비교하면 표 1과 같은 결과를 얻을 수 있게 된다 (표 1 참조).

<표 1> 대상 메타데이터 스킴 확인 사례

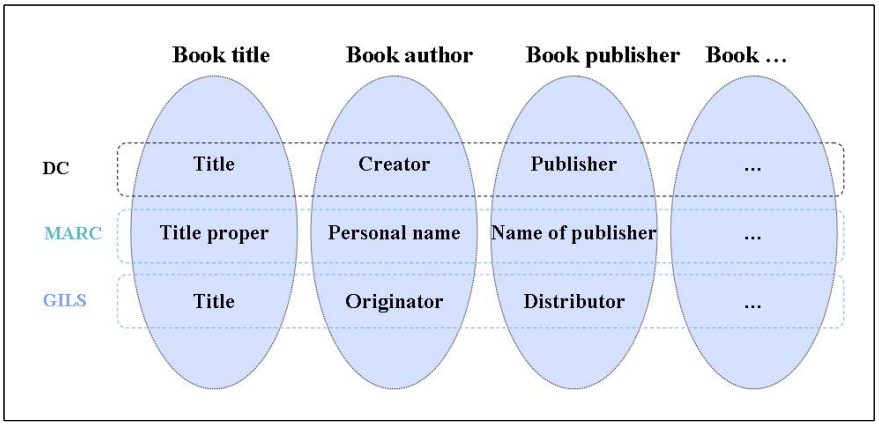
	DC	MARC	GILS
도메인 이름	전자정보자원 기술	도서관자원 캐덜로깅	정보자원 접근
필드 수	15	수 백	10(79까지 사용 가능)
사례 제시	없음	있음	없음

공통 데이터요소개념의 발견 단계에서는 먼저 전 단계에서 정해진 대상 메타데이터 중 하나의 메타데이터 셋을 기준 메타데이터로 정한다. 기준 메타데이터로는 고수준의 메타데이터 또는 단순한 구조를 가진 메타데이터가 적합하다. 이어서 기준 메타데이터를 분석하여 정보 객체(object)와 객체의 속성을 찾아내어 "ISO/IEC 11179" 표준에서 제시하고 있는 레지스트리 메타모델에 따라 데이터요소의 개념을 찾아낸다. 만일 기준 메타데이터 셋을 정하지 못한 경우는 해당 도메인을 분석하여 하향식(top down)으로 새롭게 데이터요소의 개념을 도출해야 한다. 기준 메타데이터에 의존하지 않고 하향식으로 데이터요소의 개념을 새롭게 도출하는 것은 해당 도메인과 메타데이터에 대한 전문적 지식을 요구하는 것이므로 해당 도메인의 전문가와 메타데이터 전문가가 공동으로 작업을 하는 것이 매우 중요하다. 그림 3은 공통 데이터요소개념을 발견하기 위한 온톨로지 분석 방법을 도서 목록을 대상으로 적용한 사례이다.



<그림 3> 공통 데이터요소개념의 발견을 위한 온톨로지 분석 사례

데이터요소개념에 따른 속성을 그룹화하는 단계에서는 전 단계에서 찾았거나 도출된 데이터요소개념에 의해 수집된 모든 데이터요소를 속성 수준에서 그룹핑 한다. 해당되는 데이터 요소 개념이 없을 경우에는 새로이 데이터요소개념을 추가적으로 도출한다. 그림 4는 속성들을 그룹화한 사례를 보여주는 것이다.



<그림 4> 속성의 그룹화 사례

매핑 테이블 작성 단계에서는 전 단계에서 그룹핑 한 속성들을 하나의 매핑 테이블로 작성을 하게 된다. 매핑 테이블에는 기준 데이터요소개념과 추천 데이터요소가 대상 메타데이터의 해당 속성과 함께 매핑되어 정리된다. 추천 데이터요소를 정할 때는 가급적 대상 메타데이터 가장 상위 개념에 있는 수준에 맞춤으로써 자동적 의미 호환이 가능하도록 한다. 이 때 각 속성별로 추천 데이터요소와의 의미적 일치 정도와 관계성을 표시할 경우 향후 진행되게 될 의미적 상호 호환 작업에 많은 도움을 줄 수 있게 된다. 표 2는 더블린코어, 마크, GILS를 대상으로 매핑 테이블을 작성한 사례이다.

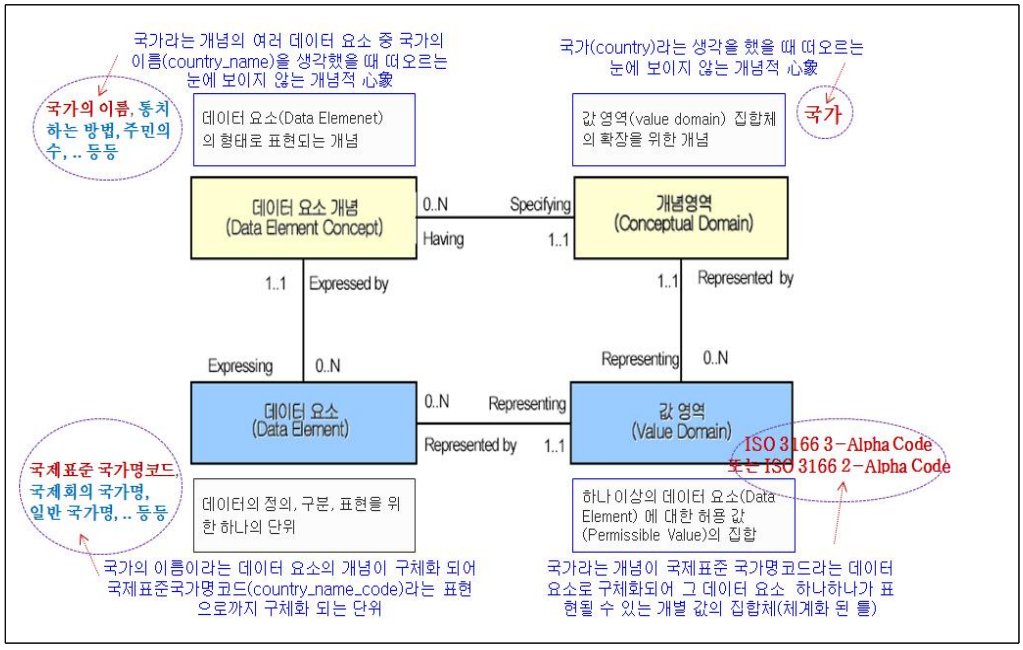
<표 2> 매핑 테이블 작성 사례

공통 (데이터요소개념)	더블린코어 (속성)	마크 (속성)	GILS (속성)	추천 (데이터요소)
Book title	Title	Title proper (L/lower)	Title	Book title
Book author	Creator (T/preferred)	Personal name (D/generic)	Originator (T/preferred)	Book author name Book author identifier . .
Book publisher	Publisher	Name of Publisher (N/order)	Distributor (T/preffered)	Book publisher name Book publisher identifier . .
---	---	---	---	---

(L: Level, D: Domain, T: Term, N: Naming rule)

3.2 데이터요소개념과 속성의 의미적 관계성 분석

의미적 메타데이터 매핑 프로세스의 두 번째 단계인 데이터요소개념 발견은 “ISO/IEC 11179-3 Registry and basic attributes”에서 제시하고 있는 메타모델을 이해하여 적용하는 것이 중요하며 (고영만, 서태설 2005, 102), 사례를 적용하여 설명한 메타모델의 개념도는 그림 5와 같다.



<그림 5> ISO/IEC 11179 Registry Metamodel의 적용 사례 개념도

의미적 메타데이터 매핑 프로세스의 네 번째 단계인 매핑 테이블 작성 사례에서 나타나는 괄호 내의 “L, D, T, N”은 의미적 매핑의 관계 유형을 표시하는 기준 척도를 의미한다. “L”은 개념의 수준 차이(level difference)를 말하는 것으로서 추천 데이터 요소를 기준으로 상위어(upper term)인지 하위어(lower term)인지를 나타내는 표시이며, “D”는 도메인의 차이(domain difference)를 말하는 것으로서 도메인의 분야나 관점의 차이를 표시한다. “T”는 사용 용어의 차이(term difference)를 나타내는 것으로 동의어(synonym)인지 유사어(preferred term)인지 반대어(antonym)인지를 표시하는 것이며, “N”은 명명 규칙의 차이(naming rule difference)를 나타내는 것으로 단어 순서(order)나 표현 방법(representation)이 다른 것을 표시하는 것이다. 표시가 없는 것은 공통 데이터 요소 개념과 일치하는 경우이다.

4. 의미적 메타데이터 매핑 적용 사례

의미적 메타데이터 매핑 프로세스의 적용을 위해 국내에서 개발된 메타데이터 표준을 조사한 결과 대부분의 메타데이터가 기관 수준에서 자체적으로 개발되어 사용되고 있는 것으로 나타났다. 최소한 단체 표준 수준에서 개발되고 또 동일 도메인에서 비교가 가능한 메타데이터로는 과학기술정보표준화위원회의 과학기술인력정보 메타데이터와 산업기술정보표준개발사업의 일환으로 개발된 산업기술인력정보 메타데이터가 있는 것으로 나타났다 (부록 1, 2 참조). 따라서 본 연구에서는 조사된 두 개의 인력정보 메타데이터를 대상으로 의미적 메타데이터 매핑 프로세스를 적용하여 매핑 테이블이 구성되는 과정을 제시하고자 한다.

4.1 대상 메터데이터 스킴 확인

과학기술인력 메타데이터는 2005년 과학기술정보화표준위원회에서 제정되어 한국과학기술정보연구원의 과학기술정보종합시스템에서 사용되고 있으며, 산업기술인력 메타데이터는 2005년 한국과학기술정보연구원에서 수행한 산업기술정보표준개발사업의 일환으로 개발되어 현재 산업기술정보데이터베이스에 적용되고 있다. 과학기술인력 메타데이터와 산업기술인력 메타데이터의 항목 수는 각각 95개와 70개, 기본 항목 수는 15개와 13개이며 (부록 1, 2 참조), 표 3은 두 메타데이터의 스킴을 비교한 결과를 보여준다 (표 3 참조).

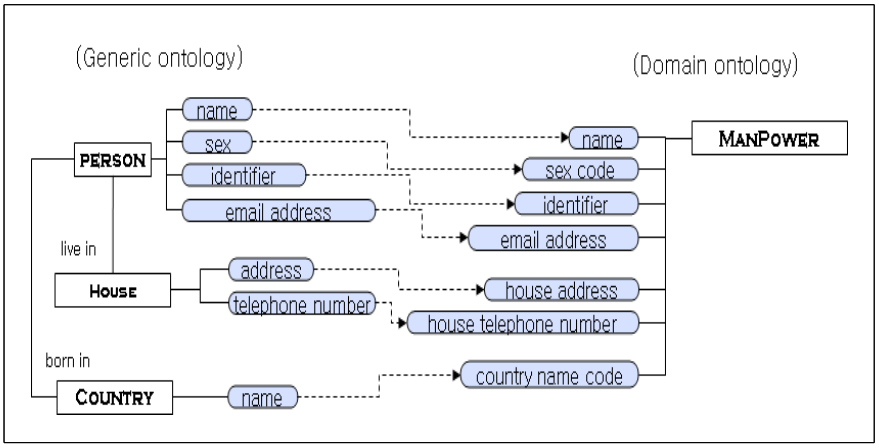
<표 3> 대상 메타데이터 확인

	과학기술인력 메타데이터	산업기술인력 메타데이터
사용 도메인	과학기술정보종합시스템	산업기술정보 데이터베이스
대상 객체	과학기술인력	산업기술인력
전체 항목 수 (기본정보 항목 수)	95개 (15개)	70개 (13개)
샘플 데이터	있음	없음
비고	과학기술정보표준화위원회에서 2005년 작성 및 제정	한국과학기술정보연구원에서 2005년 작성

4.2 공통 데이터요소개념의 발견

과학기술인력 메타데이터와 산업기술인력 메타데이터는 구조와 수준이 매우 유사하기 때문에 기준 메타데이터를 정하는 것이 매우 어려운 경우이므로 개발자의 전문성에 의

존하여 데이터요소개념을 찾아낸 다음 이들의 속성을 그룹화하는 하향식 방법을 적용하는 것이 필요하다. 이때 데이터요소개념을 효율적으로 찾아내기 위해서는 해당 메타데이터에 대한 온톨로지 분석이 많은 도움을 주게된다 (그림 6 참조).



<그림 6> 공통 데이터요소개념 발견을 위한 온톨로지 분석

온톨로지 분석에 의한 하향식의 공통 데이터요소개념 도출 과정을 통해 “인력”이라는 객체에 대한 속성으로 “이름, 식별번호, 성별, 국가, 주소, 전화번호, 이메일 주소”가 추출되었다. 추출된 7개의 속성은 우리의 심상에 “인력 이름, 인력 식별번호, 인력 성별, 인력 국가, 인력 주소, 인력 전화번호, 인력 이메일 주소”로 생각되는 개념들의 속성에 해당하는 것이다 (표 4 참조).

<표 4> 공통 데이터요소개념의 발견

	분석 결과
객체 (objects)	인력
속성 (properties)	이름, 식별번호, 성별, 국가, 주소, 전화번호, 이메일주소
데이터요소개념 (DECs)	인력 이름, 인력 식별번호, 인력 성별, 인력 국가, 인력 주소, 인력 전화번호, 인력 이메일주소

4.3 데이터 요소 개념의 속성 그룹화

하향식으로 도출된 7개의 데이터요소개념은 15개의 과학기술인력 메타데이터 기본항목과 13개의 산업기술인력 메타데이터 기본항목과는 독립적으로 설정된 개념들이다. 7개의 데이터요소개념이 두 메타데이터의 기본 항목에는 없는 개념일 수도 있으며, 반대로 두 메타데이터의 기본항목에는 있으나 추출된 7개의 데이터요소개념에는 없는 경우

도 있을 수 있다. 따라서 7개의 데이터요소개념과 두 메타데이터의 기본항목에 대한 비교가 필요하며, 비교 결과 두 개의 메타데이터 기본항목에는 있으나 추출된 7개의 데이터요소개념에는 없는 인력 출생날짜와 인력 홈페이지주소를 추가하는 것이 필요하게 되었다 (부록 1, 2 참조). 이러한 비교 과정을 통해 두 개의 데이터요소개념이 새롭게 추가되어 데이터요소개념의 수는 모두 9개로 구성되었다 (표 5 참조).

<표 5> 데이터요소개념에 따른 속성의 그룹화

공통 데이터요소개념	과학기술인력 메타데이터	산업기술인력 메타데이터
인력 이름	국문이름, 영문이름1, 영문이름2, 한문이름	한글이름, 한자이름, 영문이름
인력 식별번호	주민등록번호, 연구자번호	식별번호구분, 식별번호
인력 출생날자(추가)	생년월일	-
인력 성별	성별	성별
인력 국가	국적	국적
인력 주소	우편번호, 주소	자택우편번호, 자택주소
인력 전화번호	전화번호, 휴대전화번호	자택전화번호, 휴대전화번호
인력 이메일주소	전자우편	개인이메일
인력 홈페이지주소(추가)	홈페이지	개인홈페이지

4.4 매핑 테이블 작성

마지막 단계에서는 속성 그룹화 결과를 기반으로 도메인 전문가에 의해서 수작업으로 매핑을 수행하게 된다. 매핑 테이블에 의해 작성된 속성들의 의미적 관계를 관계유형별로 보면 가장 많은 것이 상하위어 관계(L)이고 다음으로 용어차이(T)인 것으로 나타났으며, 명명 규칙의 차이(N)도 일부 존재하였으나 도메인이나 관점의 차이(D)는 없는 것으로 나타났다. 명명규칙과 도메인의 차이가 거의 없는 현상은 인력에 관한 기본 정보만을 대상으로 데이터요소의 개념을 분석하였기 때문이라 할 수 있다.

매핑의 성립 측면에서는 상호 매핑이 되지 않는 항목(비호환 항목)이 세 개, 자동적 상호 의미 호환이 가능한 항목(호환 항목)이 다섯 개, 자동적 상호 의미 호환이 부분적으로 가능한 항목이 여덟 개로 나타났다 (표 6의 음영 표시 부분 참조). 부분호환 항목의 경우에도 추천 데이터 요소를 기준으로 해서 볼 때는 호환 항목으로 분류할 수 있다. 예를 들면 과학기술인력 메타데이터의 '주소'와 산업기술인력 메타데이터의 '자택주소'는 모두 '인력 주소'라는 데이터 요소 측면에서 의미적으로 상충되지 않기 때문에 자

동 호환할 경우 문제없이 완전 호환이 가능한 것이다. 따라서 전체 16개 항목 가운데서 13개의 항목은 추천 데이터 요소를 기준으로 볼 때 완전 호환이 가능하다 (표 6 참조).

<표 6> 매핑 테이블

공통 데이터요소개념	과학기술인력 메타데이터 속성	산업기술인력 메타데이터 속성	추천 데이터 요소
인력 이름	국문이름	한글이름 (T:syn)	인력 국문이름
	영문이름1 (L:lo,N:rep)	영문이름 (L:lo)	인력 영문이름
	영문이름2 (L:lo,N:rep)		
	한문이름	한자이름 (T:syn)	인력 한문이름
인력 식별번호		식별번호구분	인력 식별번호구분 코드
	주민등록번호(L:lo)	식별번호	인력 식별번호
	연구자번호		인력 연구자번호
인력 출생날자	생년월일		인력 생년월일
인력 성별	성별	성별	인력 성별 코드
인력 국가	국적(T:syn)	국적 (T:syn)	인력 국적국가 코드
인력 주소	우편번호	*자택우편번호 (L:lo)*	인력 우편번호 코드
	주소	*자택주소 (L:lo)*	인력 주소
인력 전화번호	전화번호	*자택전화번호 (L:lo)*	인력 전화번호
	휴대전화번호	휴대전화번호	인력 휴대전화번호
인력 이메일주소	전자우편	*개인이메일 (L:lo)*	인력 이메일
인력 홈페이지주소	홈페이지	*개인홈페이지 (L:lo)*	인력 홈페이지

L: 상하위어(up: upper term/lo: lower term), D: 도메인 관점, T: 용어(syn: synonym/ant: antonym/pre: preferred term), N: 명명 규칙(ord: order/rep: representation)

결과적으로 "의미적 메타데이터 매핑"프로세스를 적용할 경우 항목간 매핑에 중점을 두었던 기존의 단순한 메타데이터 크로스워크에서는 제대로 드러나지 않았던 의미 호환의 성격을 보다 확실히 보여주는 것이며, 또한 크로스워크 방식에 의한 메타데이터의 상호운영 효율을 대폭 향상시킬 수 있는 가능성을 보여주는 것이라 할 수 있다.

5. 결론 및 전망

본 연구에서는 메타데이터 간의 의미 호환을 극대화하기 위한 방법으로 스위칭 어크로스 방식에 기반한 "의미적 메타데이터 매핑 프로세스"를 제안하였다. 스위칭 어크로스 방식은 메타데이터 크로스워크의 한 유형으로 데이터개념요소의 분석과 도출에 온톨로지 개념을 도입할 경우 매핑의 수준을 의미적 호환까지 끌어올릴 수 있는 방법이다. 본 연구에서 제안한 "의미적 메타데이터 매핑 프로세스는" 대상 메타데이터의 스킴 확인, 공통 데이터 요소개념의

확인, 데이터요소개념의 속성 그룹화, 매핑 테이블 작성의 네 단계로 구성된다. 또한 네 번째 단계에서는 매핑되는 개념 들의 관계유형을 개념의 수준 차이(L), 도메인 차이(D), 사용 용어 차이(T), 명명 규칙 차이(N) 등으로 구분하는 기준 척도를 제시하였다.

본 연구에서 제안된 “의미적 메타데이터 매핑 프로세스”를 동일한 도메인에서 개발된 두 개의 인력정보 메타데이터 단체 표준에 적용해서 매핑을 시도한 결과 대체로 무리없이 적용되는 것으로 나타났으며, 기존의 단순한 메타데이터 크로스워크에서는 제대로 드러나지 않았던 의미 호환의 성격을 보다 확실히 보여주었다. 따라서 본 연구에서 제안한 “의미적 메타데이터 매핑 프로세스”를 보다 정교하게 개발할 경우 기존의 단순한 항목 간 매핑 방식의 메타데이터 크로스워크에서는 파악할 수 없었던 의미 간 호환의 유형을 보다 명확히 보여주게 되어 크로스워크를 통한 메타데이터 상호운영 효율을 향상시키는 데 많은 도움이 될 수 있을 것으로 기대된다.

그렇지만 본 연구에서 제안된 “의미적 메타데이터 매핑 프로세스”가 실제에서 무리없이 사용되기 위해서는 다양한 메타데이터에 확대 적용함으로써 프로세스의 완성도를 높이는 작업이 이루어져야 할 것이다. 이를 통해 매핑 프로세스를 최대한 자동화 할 뿐만 아니라 매핑의 관계유형을 보다 세분화하고 정형화하는 작업이 축적되어야 하며, 데이터 요소 외에 데이터 값의 상호 매핑 방안에 관한 연구도 추가로 수행되어야 할 것이다. 특히 자동 매핑의 비율을 최대화하기 위해서는 “의미적 메타데이터 매핑 프로세스”를 기반으로 하여 개발된 다양한 메타데이터 표준간의 크로스워크를 공개함으로써 많은 사람들이 관련 연구와 개발 결과를 실제에 활용할 수 있도록 하는 것이 매우 중요하다.

참 고 문 헌

고영만. 2005. 온톨로지와 웹 온톨로지.『메타데이터 표준화 포럼 제1회 워크숍 자료집 - 메타데이터와 온톨로지』. 2005.3.9, 서울: 한국과학기술정보연구원.

고영만, 서태설. 2005. 온톨로지 기반 메타데이터 명명규칙에 관한 연구. 『정보관리학회지』, 22(4): 97-109.

과학기술정보표준화위원회. 2005. 과학기술 인력정보를 위한 메타데이터. STI-S.2005-3.03.

서태설, Pham, D. T. 2007. 데이터의 의미적 상호운용성 확보를 위한 데이터 모델링 프로세스: EDM 가공에의 적용. 『정보관리연구』, 37(1): 59-73.

한국과학기술정보연구원. 2005. 국가산업기술정보 메타데이터 시범구축 사업: 인력 정보 메타데이터 표준안. 산업자원부 기술표준원

한국데이터베이스진흥센터. 2006. 데이터베이스 연계를 위한 통합 메타모델 개발(1차년도). 정보통신부.

홍성화, 서태설. 2004. 분류체계 일치를 통한 과학기술정보 상호 교환 방법에 관한 기초 연구. 『정보관리연구』, 35(3): 109-123.

Burgman, Michael K. 2006. "Models of Semantic Interoperability", Blog. [cited 2006. 12. 14]. <http://www.mkbergman.com/?cat=16>.

Chan, L. M., Zeng, M. L. 2006. "Metadata Interoperability and Standardization: A Study of Methodology (Part I) Achieving Interoperability at the Schema Level", D-Lib Magazine, 12(6), [cited 2007. 10. 8]. <http://www.dlib.org/dlib/june06/chan/06chan.html>
<http://www.slis.kent.edu/~mzeng/metadata/crosswalks.htm>

Godby, C. J., Young, J. A., Childress, E. 2004. "A Repository of Metadata Crosswalks", D-Lib Magazine, 10(12), [cited 2007. 10. 8].
<http://www.dlib.org/dlib/december04/godby/12godby.html>

Fowler, J., Perry, B., Nodine, M. and Bargmeyer, B. 1999. "Agent-Based Semantic Interoperability in InfoSleuth". SIGMOD Record, 28(1): 60-67.

ISO/IEC JTC1. 2003. ISO/IEC 11179 Information Technology - Metadata Registries

Lightle, K. S., Ridgway, J. S. 2003. "Generation of XML Records across Multiple Metadata Standards", D-Lib Magazine, 9(9), [cited 2007. 10. 8].
<http://dlib.org/dlib/september03/lightle/09lightle.html>

Pierre, M. St., LaPlant, W. P. Jr. 1998. "Issues in Crosswalking Content Metadata Standards" [cited 2007. 10. 8]
<http://www.niso.org/press/whitepapers/crsswalk.html>

부록 1 : 과학기술인력정보 메타데이터 기본 항목

속성명(영문)	데이터 타입 (길이)	필수 /조건부필수 /선택	개요 및 사례	반복	코드
주민등록번호/ 외국인등록번호 (Personal ID Number)	char(14)	필수	연구자 개인의 고유성을 구별할 수 있는 핵심 식별자. 외국 국적 연구 인력일 경우 외국인등록번호사용	불가	
연구자번호(Research ID Number)	char(20)	선택	기관별로 부여한 연구자 관리 및 식별자	불가	
국문이름(Name)	char(20)	필수	주민등록증 상의 본명 혹 실명	불가	
영문이름1 (Family name)	char(64)	필수	여권에 기재되어 있는 국제적으로 통용가능한 연구자의 성.	불가	
영문이름2 (First/Middle name)	char(32)	필수	여권에 기재되어 있는 국제적으로 통용가능한 연구자의 이름	불가	
한문이름 (Chinese name)	char(32)	선택	주민등록증상의 한문이름	불가	
성별(Sex)	char(1)	필수	남녀 구분 (예) 미상, 남, 여, 기타	불가	성별코드: ISO/IEC 5218
생년월일 (Date of Birth)	date	필수	주민등록증 및 여권 등에 공식적으로 인정되는 생년월일 (예) 2004.-03-22	불가	날짜표기: ISO 8601 YYYY-MM-DD
국적(Nationality)	char(3)	필수	연구자의 국적 기재 (예) KOR, GBR, USA	불가	국가코드: ISO 3166
우편번호 (Zip Code)	char(7)	선택	자택, 개인 사무실 등의 사적 주소의 우편번호	반복	우편번호 코드
주소(Address)	char(64)	선택	자택 등 연락 가능한 주소	반복	
전화번호(Telephone)	char(14)	필수	연구자와 연락 가능한 전화번호	반복	
휴대전화번호 (Cellular phone)	char(14)	선택	긴급 상황 시 연락 가능한 전화번호	반복	
전자우편 (E-mail Address)	char(64)	조건부필수	연구자와 가상의 공간에서 연락 가능한 이메일주소	반복	
홈페이지(URL)	char(128)	선택	연구자 개인의 홈페이지	반복	

부록 2 : 산업기술인력정보 메타데이터 기본 항목

클래스	데이터식별자	데이터요소이름	데이터 유형	데이터 유형 참조 스킴
인력	HMDE001	인력_식별번호구분_코드	NUMBER	자체구분코드부여
	HMDE002	인력_식별번호_코드	NUMBER	주민등록번호, 여권번호, 외국인등록번호
	HMDE003	인력_한글이름	STRING	
	HMDE004	인력_한자이름	STRING	
	HMDE005	인력_영문이름	STRING	
	HMDE006	인력_성별_코드	NUMBER	ISO 5218
	HMDE007	인력_국적_코드	STRING	ISO3166-1
	HMDE008	인력_자택우편번호_코드	NUMBER	행정표준코드체계
	HMDE009	인력_자택주소	STRING	
	HMDE010	인력_자택전화번호	NUMBER	
	HMDE011	인력_휴대전화번호	NUMBER	
	HMDE012	인력_개인홈페이지	STRING	
	HMDE013	인력_개인이메일_주소	STRING	