데이터 우선 순위 기반의 점진적인 데이터 통합을 위한 메타 모델 설계1)†

유상훈0* 정동원* 신동길* 서태설** 백두권*
고려대학교 소프트웨어 시스템 연구실
{shryu0, withimp, dkshin, baik}@software.korea.ac.kr
{tsseo}@kmail.kisti.re.kr

A Design of Metamodel for Progressive Data Integration Based on Data Priority

Sang-Honn Ryu0*, Dong-Kil Shin*, Dong-Won Jeong*, Tae-Seol Seo**, Doo-Kwon Baik*
*Dept. of Computer and Engineering, Korea University
**Knowledge Information Standardization Lab, KISTI

요 약

데이터베이스를 통합하기 위한 많은 연구들이 진행되어 왔지만, 통합하고자 하는 모든 데이터들을 고려함으로써 초기 비용과 시간에 대한 오버헤드로 인해 비효율적이며 현실적으로 불가능한 경우가 발생하게 된다. 이 논문에서는 이러한 문제점을 개선하여 점진적인 통합을 위한 개념적인 통합 방법론을 제안하고, 제안된 방법론을 한국과학기술정보연구원에서 보유하고 있는 데이터베이스 통합에 적용하기 위한 통합 메타 모델을 설계한다. 또한 다른 기관 또는 다른 포맷들과의 상호운용성을 향상시키기 위하여 해당 분야의 국제 표준 또는 사실 표준들을 고려하여 통합 메타 모델을 설계하였다.

1. 연구 배경

지금까지 데이터베이스 통합을 위한 많은 연구들이 진행되어 왔으며, 데이터베이스 통합은 분산 구축 및 관리되는 데이터들에 대한 일관성과 표준화된 관리를 통해, 양질의 서비스를 제공하는데 그 일차적인 목적이 있다[1,2,3,4,5,6]. 또한 향후, 관리 측면에서 비용과 시간이 절약되고 상호운용성면에서 다른 분야의 데이터와의 관계를 통한 보다 나은 서비스를 제공할 수 있다.

한국과학기술정보연구원에는 국내에서 과학기술정보와 관련한 많은 분야의 데이터베이스를 보유하고 있다. 그러나 각 분야에만을 고려하여 설계되고, 유사한 분야일 경우에도 서로 다른 스키마를 정의하여 사용함으로써 데이터 공유와 교환이 불가능하다. 또한 비정규화된 설계로 인한 중복성 문제와 일관성 있는 관리가 어렵다. 즉, 이질성의 데이터베이스들의 통합에서 일관성을 유지하기 어렵고 상호운영을 하기 힘들다는 문제점을 가지고 있다.

이러한 문제점을 해결하고 통합된 서비스를 제공하기 위하여 통합된 데이터 모델이 요구된다. 그러나 현재까지 연구되어 온 통합 방법은 모든 데이터베이스들에 대한 통합방법으로서, 초기 비용과 시간이 과다하게 소요되고 현실적으로 불가능한 경우도 발생한다.

본 논문에서 이러한 기존 방법의 문제점을 개선하기 위한 데이터 우선 순위 기반의 점진적인 데이터베이스 통합 방법 제안한다. 특히, 실질적인 적용을 위한 통합 메타 모델 설계에 초점을 둔다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존 통합방법론과 특성에 대하여 기술하고, 3장에서 이 논문에서 제안하는 점진적인 통합 메카니즘과 실제 적용하여 설계한 통합 메타 모델에 대하여 간략하게 기술한다. 제 4장에서는 시스템 구조에 대하여 기술하고 5장에서 결론과 향후 연구 방향에 대하여 기술한다.

† 본 연구는 한국과학기술정보연구원의 지원으로 수행되었음

2. 관련 연구

대부분의 데이터베이스 통합 접근법은 주로 상향식(bottom-up) 통합방법이다. 상향식 통합방법은 통합을 하고자 원하는 데이터베이스의 분석과 설계를 통해 공통 스키마를 추출하고 정의함으로써, 지역 스키마 사이의 다른 점들을 포함하는 하나의 전역 스키마를 설계하는 방법이 있다. 전역 스키마 또는 연방 통합 방법 등이 여기에 해당한다[1,2]. 그 외에 다른 통합방법으로는 분산 객체 통합방법(distributed object approach), 중개자 기반 통합방법(mediator-based approach), 추론 기반 통합방법(case-based reasoning-based approach) 등이 있다[3,4,5,6]. 이러한 접근방법들은 기본적으로 온톨로지 기반 통합 방법론으로써 분류되며, 실제 데이터베이스가 갱신될 때마다 계속적으로 온톨로지는 변화되고 갱신되기 때문에, 관리를 위해 많은 비용이 든다. 또한 온톨로지에 대한 표준화가 어렵기 때문에 통합 사이클이 계속적으로 반복되게 된다.

또 다른 형태의 통합방법이 하향식(top-down) 통합방법이다. 하향식 접근방법은 먼저 공통 스키마를 추출하고 정의함으로써 데이터베이스를 통합한다. 따라서 이 통합방법은 새로운 데이터베이스를 구축하기 위해 표준화된 지침을 제공한다. 이 통합방법의 표준화된 지침은 데이터를 관리하는데 여러 장점을 가진다. 갱신에 따른 비용이 감소되고 데이터 값의 의미에 대한 표준화된 스키마와 정의로 일관성 있게 관리할 수 있다. 일반적으로 메타데이터 포맷이라 불리는 이러한 지침은 각 분야별로 연구되고 있으며, 이 논문에서 대상으로 하는 문헌 정보를 위한 표준 메타 포맷에는 MARC[8], Dublin Core, ONIX, RDF[7] 등이 있다. 그러나 이러한 포맷은 동적 메타데이터 포맷 관리를 위한 방법이 지원되지 않는다는 문제점이 있다. 이러한 문제를 해결하기 위해 정보 기술 분야에서는 ISO(표준화

국제 기구)와 IEC(국제 전자기술 위원회)는 ISO/IEC JTC1이라는 공동 기술 위원회를 설립하였다. 특히 ISO/IEC 11179는 공유된 데이터 환경을 적은 시간과 적은 노력으로 생성하기 위한 목적을 지니며, 실세계에서 모든 사물은 인식 가능한 이름처럼 색깔, 크기, 식별자와 같은 속성을 가진다. 속성은 데이터로써 표현된다[7,8].

3. 통합 메타데이터 모델

3.1 대상 도메인

이 연구에서는 많은 데이터베이스들 중에서 해당 분야에서 데이터 우선 순위가 높은 데이터베이스들을 선정하여 통합 메타데이터 모델을 설계한다. 현재 한국과학기술정보연구원에서 보유하고 있는 데이터베이스 중에서 과학기술 관련 문헌정보 데이터베이스 중심으로 서비스를 제공하고 있다. 따라서 이 논문에서는 과학기술 지식정보 관련 문헌 정보 데이터베이스들은 대상 도메인으로 선정하여 설계하였다.

대상 데이터베이스들로는 DIGS(전분야 국내 정기 간행물 기사 색인), DIMD(전분야 국내 석박사 학위 논문 목록), INFO(문헌 벙보 분야 학술지 서지 정보), THESIS(석박사 학위 논문), KRIST(연구보고서), BIST(국내외 과학 기술 문헌 정보), SATURN(과학 기술 전문 정보), SOCIETY(국내 학회 논문 정보)의 데이터베이스를 선정하여 이들 간의 데이터를 공유하고 교환할 수 있도록 통합 메타 모델을 설계하였다.

3.2 개념적 통합 메타데이터 모델

앞서 기술하였듯이, 논문에서 제안한 통합 방법의 목적은 상호운용성과 데이터 통합과 및 점진적 통합 메커니즘들을 제공하고자 하는 것이다. 이를 위해서는 각 도메인의 표준과 한국과학기술정보연구원에서 사용하는 데이터 요소와의 결합된 모델이 요구되며 점진적인 통합을 위한 ISO/IEC 11179의 MDR과 결합된 모델이 우선적으로 요구된다. 다음 그림은 이와 같은 메커니즘을 도식한 것이다.

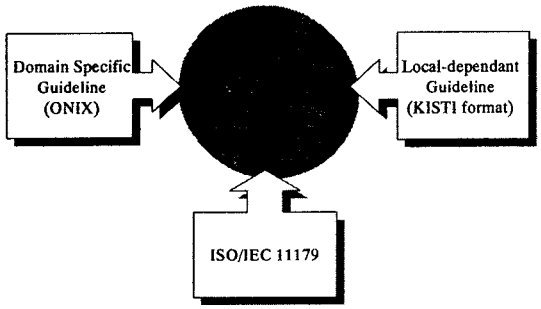

그림 1. 개념적 통합 메타데이터 모델

이 논문의 목적은 한국과학기술정보연구원에 있는 데이터베이스 중에 문헌정보 데이터베이스를 대상으로 하여 제안한 통합 방법을 적용하는데 있다. 이를 위해서 문헌정보 분야에서 대표적인 지침으로 사용되고 있는 ONIX와 한국과학기술정보연구원의 대상 데이터베이스, 그리고 ISO/IEC 11179와의 통합 모델이 요구된다.

3.3 논리적 통합 메타데이터 모델

여기에서는 데이터 우선 순위 기반의 통합 방법을 문헌정보 데이터베이스에 적용하고자 한다. 각각 요구되는 메타모델들에 대하여 기술한다.

3.3.1 데이터 요소의 메타 모델

다음 그림은 한국과학기술정보연구원에서 제공한 7개의 데이터베이스들을 의미별로 정규화시킨 것이다.

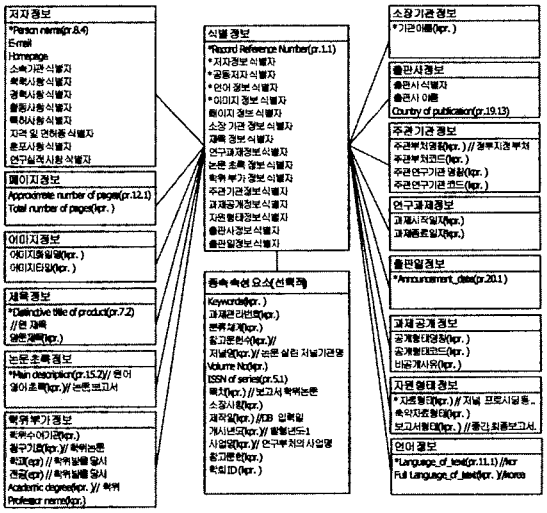

그림 2. 데이터 요소의 메타모델

데이터베이스들 간의 서로 다른 타입과 이름으로 사용되고 있는 스키마들을 분석하여 동일한 의미를 가지는 데이터들을 추출한다. 그리고 동일한 의미의 데이터들을 매핑시키고 ONIX에서 동일한 데이터 요소를 찾아 매핑시킨다. 모든 스키마들을 분석하였다면 7개의 모든 데이터베이스에서 필요한 식별 정보와 각 데이터베이스에서 제한적으로 사용되는 종속 속성 요소(선택적)를 나뉘게 된다. 필수적으로 사용되어야 하는 식별정보를 살펴보면 저자 정보, 제목 정보, 논문 초록 정보, 출판사 정보, 소장 기관 정보 등과 같이 꼭 필요한 정보를 담은 데이터들에 대한 식별자를 가지고 있고 이 식별자를 통하여 참조되는 테이블은 그에 대한 더욱 상세한 데이터들을 가지고 있다. 이렇게 정규화를 시켜 데이터베이스를 통합한다면 데이터 중복을 막을 수 있으며 구조를 균일화 할 수 있다.

3.3.2 데이터 요소의 메타 모델

다음 그림은 데이터 요소 식별자와 사상 영역 식별자가 데이터베이스를 접근하기 위한 메타모델을 보여준다.

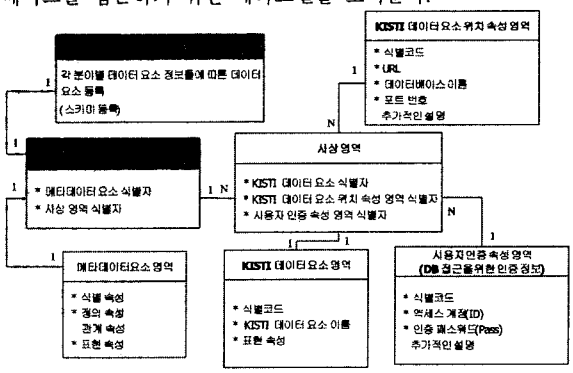

그림 3. 데이터 요소의 메타모델

데이터 요소 Concept 식별 영역에서 통합된 메타데이터 요소 식별자를 통해 MDR로부터 그에 해당하는 각 데이터베이스의 데이터 요소에 대한 정보를 얻는다. 그리고 통합된 메타데이터 요소와 KSITI 데이터 요소를 사상시킨 사상영역에서 KISTI 데이터 요소 위치 속성 식별자를 통해 url, 데이터베이스 이름, 포트 번호를 얻어서 KISTI 데이터베이스를 접근한다. 이때 인증 절차가 필요한데, 사용자 인증 속성 영역에서 계정과 패스워드로 인증을 한다. 그리고 나서 모든 각 데이터베이스로부터 데이터를 가지고 와서 사용자에게 보여준다.

3.3.3 메타데이터 등록 및 관리 메타 모델

다음 그림은 도메인 전문가가 필요로 하는 메타데이터 등록과 등록된 메타데이터를 관리하기 위한 메타 모델을 보여준다.

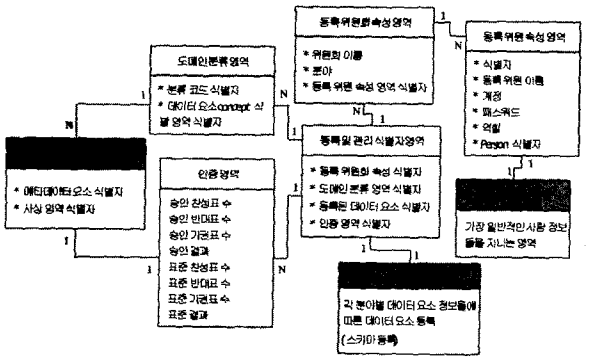

그림 4. 메타데이터 등록 및 관리 메타 모델

전문가로써 인증이 되면 도메인 전문가가 데이터 요소 Concept 식별 영역을 통해 그 영역별로 쓰이는 MDR의 데이터 요소에 대한 정의와 의미를 파악하고 도메인 분류 영역에서 자신이 등록하고자 하는 데이터 요소를 분류하여 등록을 한다. 그러면 등록 위원회 속성 영역에서 해당 도메인 영역의 등록 위원회가 소집되어 등록된 데이터 요소를 등록할 것인지를 면밀하게 검토하고, 인증 영역에서 표결로 승인과 표준 절차를 거쳐서 데이터 요소와 일치하게 만들어졌다면 MDR에 추가한다. 등록 위원회로부터 승인을 받지 못할 경우에는 등록을 원하는 전문가와 접촉을 하여 조정을 하게 된다.

3.3 논리적 통합 메타데이터 모델

데이터 요소의 메타모델, 통합 검색을 위한 메타모델, 메타데이터 등록 및 관리 메타 모델 간의 상호 관계성을 보여주고 있다.

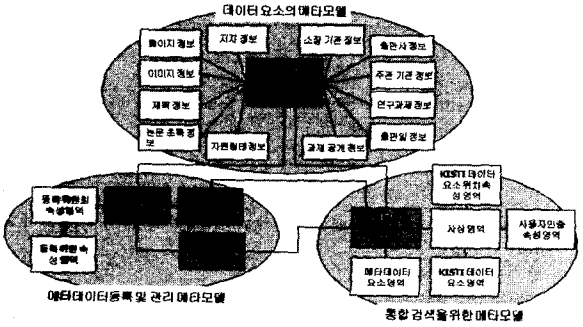

그림 5. 통합 메타모델

데이터 요소의 메타모델은 데이터베이스들의 공통 요소를 추출하여 메타데이터 요소를 만들었고 이 메타데이터 요소 식별 정보들은 메타모델의 등록 및 관리 식별자 영역과 연결되어 메타데이터 요소들을 관리하고 검색할 수 있는 관계성을 보여 준다. 등록 및 관리 식별자 영역에서 메타데이터를 등록하였다면 등록 위원회 속성 영역에서 그 분야의 위원회가 심사를 한다. 등록된 메타데이터 요소가 인증되면 MDR에 추가시켜 사용할 수 있도록 한다.

그리고 메타데이터 요소 식별정보는 통합 검색을 위한 메타모델의 데이터 요소 Concept 식별자 영역과 연결되어 사상 영역을 통해 KISTI 데이터 요소 위치 속성 영역에서 데이터의 위치, 계정, 패스워드를 얻어서 사용자 인증 속성 영역에서 사용자를 인증해 준다.

4. 결론 및 향후 연구

문헌 정보쪽에서 구축된 데이터베이스들이 각자 자신의 분야에 맞게 데이터를 구성하고 있으므로 그들 사이에 데이터를 공유하지 못한다. 따라서 그 데이터베이스에 종속적인 데이터를 검색해야 하며 다른 데이터베이스에서도 동일한 검색을 해야 한다. 이러한 불편함을 제거하고 사용자에게 양질의 서비스를 제공하며 분산된 데이터베이스들의 표준화된 관리를 위해 통합 데이터베이스가 필요하다.

그러나 현실적으로 통합하고자 하는 모든 데이터베이스를 통합하는 것은 초기 비용과 시간의 제약이 있었다. 이를 해결하기 위한 방법으로 본 논문에서 데이터 우선 순위 기반의 점진적인 데이터 통합을 제안하였다. 이 접근법은 데이터가 중요한 것부터 데이터베이스를 점진적으로 통합함으로써 시간을 단축하고 초기 비용을 감소시킬 수 있다. 그리고 점진적으로 데이터베이스를 추가로 통합함으로 결국에는 모든 데이터베이스에서 동일한 의미를 가지는 데이터를 통합하여 각 데이터베이스에서 검색할 수 있다. 그리고 이후에 구축되는 데이터베이스들은 모두 통합 데이터 모델을 따르기 때문에 데이터베이스들 간의 데이터가 일관성 있는 의미를 가지게 되며 이기종의 시스템에서 공유와 교환이 가능하게 된다. 또한 다른 분야의 데이터베이스들과 상호운용성을 증가시킬 수 있다. 향후에는 통합 모델을 기반으로 보다 세부적인 설계와 시스템 구현을 통해 문제점 및 개선방안을 도출하고자 한다.

5. 참고문헌

[1] Ram, S., Special issue on heterogeneous distributed database systems, IEEE Computer Magazine, 24, 12 (December 1991).

[2] Ahmed and et al, The Pegasus heterogeneous multidatabase system, IEEE Computer, 24, 12 (1991).

[3] Panti, M., Spalazzi, L., Giretti, A, A Case-Based Approach to information Integration, Proceedings of the 26th VLDB Conference (2000).

[4] Manola, F. and et al, Distributed object management, International Journal of Intelligent and Cooperative Information Systems, 1, 1 (March 1992).

[5] Ozsu, T., Dayal, U. and Valduriez, P., Distributed Object Management, Morgan Kaufmann, San Mateo, CA (1993).

[6] Wiederhold, G., Mediators in the Architecture of Future Information Systems, IEE Computer Magazine, 25 (March 1992), 38-49.

[7] http://www.w3.org/RDF/

[8] http://www.jtc1sc32.org/