

MDR과 온톨로지를 결합한 3계층 정보 통합 시스템

백 두 권[†]·최 요 한^{††}·박 성 공^{†††}·이 정 옥^{††††}·정 동 원^{††}

요 약

한 도메인 내에서 다양한 데이터베이스의 데이터를 공유하고 표준화하기 위해 MDR(Metadata Registry)을 이용하여 정보를 통합할 수 있다. 그러나 MDR을 구축하는 조직간 데이터요소 표현의 불일치 때문에 MDR간 광역적인 정보를 통합하는데 어려움이 있다. 또한 웹과 같은 다양한 데이터베이스가 존재하는 환경에서 통합된 정보를 검색하고자 하는 사용자는 각각의 데이터베이스 스키마 정보를 확보하기엔 한계가 있다. 따라서 본 논문에서는 MDR과 온톨로지(Ontology)를 결합한 3계층 정보 통합 시스템을 제안한다. MDR간 데이터요소의 관계를 사상 시키고 표현의 불일치를 해결하기 위해 MDR의 표준성기능과 온톨로지의 개념과 관계기능을 결합한 정보 통합 모델을 정의하고 에이전트 기술을 적용한 계층적이고 독립적인 정보 통합 아키텍처를 제안한다. 온톨로지는 사용자의 질의에서 개념을 추출하기 위한 의미망(semantic network)의 역할과 MDR간의 데이터요소 관계를 설정하기 위한 기능으로 적용되었다. MDR과 지식베이스(Knowledge Base)는 데이터요소간 표현 불일치를 해결하기 위해 적용하였다. 이러한 핵심요소를 고려하여 제안된 아키텍처를 사용하여 MDR과 온톨로지를 결합한 3계층 정보 통합 시스템을 구현하였다.

A 3-Layered Information Integration System based on MDRs and Ontology

Doo-Kwon Baik[†] · YoHan Choi^{††} · SungKong Park^{†††}
JeongOog Lee^{††††} · DongWon Jeong^{††}

ABSTRACT

To share and standardize information, especially in the database environments, MDR (Metadata Registry) can be used to integrate various heterogeneous databases within a particular domain. But due to the discrepancies of data element representation between organizations, global information integration is not so easy. And users who are searching integrated information on the Web have limitation to obtain schema information for the underlying source databases. To solve those problems, in this paper, we present a 3-layered Information Integration System (LIIS) based on MDRs and Ontology. The purpose of proposed architecture is to define information integration model, which combine both of the nature of MDRs standard specification and functionality of ontology for the concept and relation. Adopting agent technology to the proposed model plays a key role to support the hierarchical and independent information integration architecture. Ontology is used as for a role of semantic network from which it extracts concept from the user query and the establishment of relationship between MDRs for the data element. MDR and Knowledge Base are used as for the solution of discrepancies of data element representation between MDRs. Based on this architectural concept, LIIS was designed and implemented.

키워드: MDR(Metadata Registry), 온톨로지(Ontology), 에이전트(Agent), 정보 통합(Information Integration)

1. 서 론

데이터베이스 시스템은 특정 요구사항에 따라 독립적으로 생성, 저장, 사용되어 왔다. 그러나 인터넷의 발달은 사용자로 하여금 개별 데이터베이스에 구축되어 운영되던 정보를 통합하여 새로운 정보를 얻으려는 필요성을 증가시켰다. 또한 분야간 다양한 정보를 교류해야 하는 정보화 사회

의 흐름에 맞추어 단독 운영되던 데이터베이스는 상호 연관을 가지게 되었다. 이것은 정보 시스템간 특히, 데이터베이스간의 통합을 요구한다. 그러나 통합하려는 데이터베이스 시스템은 플랫폼 이질성, 의미 이질성 등을 가지게 되며 이러한 이질성을 해결하기 위해 다양한 통합 방법이 등장하게 되었다[1]. 초창기에는 개별 데이터베이스의 스키마를 전역스키마로 통합하는 방법들이 제안되었다. 그러나 전역스키마는 지역소스의 추가나 수정시 갱신이 요구되는 부담이 존재한다. 전역스키마의 정적인 접근 방법을 해결하기 위해 멀티 데이터베이스 언어를 이용한 방법이 제시되었으나 사용자로 하여금 복잡한 전역 인터페이스 사용을 요구

* 이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2000-041-E00286).

† 종신회원: 고려대학교 컴퓨터학과 교수

†† 준 회원: 고려대학교 대학원 컴퓨터학과

††† 정 회원: 건국대학교 컴퓨터·응용과학부 교수

논문접수: 2002년 8월 3일, 심사완료: 2002년 12월 31일

한다[5]. 이 후 웹과 같은 동적인 환경에서 적용 가능한 지능형 중재기, 도메인 온톨로지, 에이전트, 그리고 고수준의 언어를 이용한 방법들이 제안되었다. 하지만 사용자가 통합해야 하는 도메인에 대한 지식을 알고 있어야 하는 가정이 문제점으로 작용한다. 최근에는 온라인 사전과 시소러스 등을 이용한 Summary Schemas Model(SSM)이 제안되어 기존 멀티 데이터베이스 시스템에 대한 확장으로 의미 식별을 지원하기도 한다. 그러나 규모가 큰 멀티 데이터베이스 시스템에서는 성능의 한계를 가져온다.

본 논문에서는 사용자의 도메인에 대한 지식과 개별 데이터베이스의 스키마 정보확보에 대한 부담을 줄이고 개별 데이터베이스 시스템의 추가를 유연하게 제공하는 MDR과 온톨로지를 결합한 3계층 정보 통합 시스템(LI2S : A 3-Layered Information Integration System)을 제안하였다.

MDR(Metadata Registry)은 특정 도메인 영역에서 다양한 형태로 개발되어진 데이터베이스의 데이터들에 대한 정보 통합 및 유통을 위해 표준화된 데이터요소를 등록하여 활용하는 방법이다[2]. 그러나 이렇게 표준화 되어 있는 MDR도 특정 도메인에 맞게 구성되어 있기 때문에 MDR을 구성한 조직간 데이터요소 의미 이질성이 발생할 수 있다. 논문에서 접근하고자 하는 정보 통합의 조건은 MDR을 사용하여 정보 통합을 이룬 조직간의 이질적인 정보를 광역적인 정보로 통합하고자 할 경우이다. 온톨로지(Ontology)는 다양한 도메인의 개념과 관계를 정의하여 정보 통합에 사용될 수 있다[3]. 온톨로지는 MDR의 데이터요소들을 개념화하고 관계정보를 유지하여 사용자의 질의와 데이터요소간 스키마 이질성을 해결하기 위해 사용한다. 데이터요소간 형식, 크기, 그리고 단위 등과 같이 표현의 차이로 생기는 데이터 이질성은 지식베이스(KB : Knowledge Base)를 구축하여 해결한다. 이러한 온톨로지와 지식베이스의 개념을 검색하여 데이터요소간 사상(Mapping)과 이질성을 해결하기 위해서는 자율적이고 지능적인 에이전트가 요구되고 개별 MDR에 연결되어 있는 하위 데이터베이스들간의 사상을 위해서는 상호운용성(Inter-operability)을 지닌 에이전트를 필요로 하게 된다. 새로운 데이터베이스 소스의 추가시 확장의 유연성 측면에서 각 에이전트들이 독립적인 기능을 수행하도록 계층화 된 아키텍처를 구성하였다.

논문의 구성은 다음과 같다. 2장에서는 정보 통합 방법에 대한 관련 연구와 논문에서 사용되는 의미 이질성의 분류, 온톨로지 및 MDR에 대해 소개한다. 3장에서는 MDR과 온톨로지를 결합하여 정보를 통합하는 모델에 대한 설명과 MDR과 온톨로지의 역할에 대해서 기술한다. 4장에서는 MDR과 온톨로지를 결합한 3계층 아키텍처의 정보 통합 시스템을 제안하고 각 계층별 에이전트 구성과 핵심 구성요소 및 정보 통합 과정을 설명한다. 5장에서는 LI2S 구현에 대해 설명한다. 6장에서는 LI2S를 기존의 정보 통합 시스템과 비교

한 평가와 특징에 대해 기술한다. 마지막으로, 7장에서는 결론을 기술한다.

2. 관련 연구

멀티 데이터베이스 시스템(multidatabase system)은 상호 이질적인 지역 DBMS에 대한 통합 인터페이스를 제공하는 분산 시스템이다[4]. 멀티 데이터베이스에 대한 통합된 접근을 위해서는 의미 이질성이 해결되어야 한다. 의미 이질성은 동일한 개념의 스키마가 지역 데이터베이스 스키마에 다양하게 표현되어 있을때 발생한다. 이러한 의미 이질성을 해결하기 위해 다양한 방법이 제안되어 왔다. 초기에는 고전적 접근 방식으로 다양한 지역 데이터베이스 스키마의 차이를 해결하기 위해 하나의 전역 스키마를 구성하는데 초점을 맞추었다. Multibase, Pegasus, TSIMMIS, 그리고 Carnot 등과 같은 시스템이 존재한다[5-7]. 특히 MCC의 Carnot 프로젝트는 데이터베이스 시스템, 데이터베이스 응용, 전문가 시스템을 포함하는 다양한 시스템으로부터의 이질적 정보를 논리적으로 통합하는 시스템이다[8]. 다양한 기능으로 세분화되어 개발되었는데 이중 Semantic Service에서 모든 정보 자원의 전역적 뷰(global view)를 제공한다. 전역 스키마 멀티 데이터베이스는 사용자에게 단일의 통합된 뷰를 제공하지만 지역 데이터베이스가 새로 추가되거나 수정되면 전역 스키마의 갱신이 요구되는 단점이 있다. 그 다음으로 제안된 방법이 멀티 데이터베이스 언어이다. 멀티 데이터베이스 언어를 지원하는 시스템에서는 전역스키마를 유지하지 않으며 사전에 데이터베이스 스키마 정보에 대해 알 필요가 없다. 제안된 시스템으로는 Calida, MRDSM, 그리고 SYBASE 등이 있다[9]. 이 중 SYBASE는 다양한 데이터베이스의 통합 필요에 따라 기존의 분산된 이질적 데이터베이스 접근 기능을 확장한 Open Server를 개발하고 네트워크의 연결과 클라이언트 프로그램으로부터 요청관리를 하는 서버 네트워크 인터페이스를 제공한다. 그러나 멀티 데이터베이스 언어 시스템은 분산된 데이터베이스에 대한 투명성을 제공하지 않으며 사용자가 일관성을 유지해야 하는 단점이 있다. 다음으로 제안된 방법들은 에이전트(agent), 도메인 온톨로지(domain ontology), 지능형 중재기(intelligent mediator), 그리고 고수준 질의어와 같은 다양한 기술을 이용하여 동적인 환경에 적용하도록 제안되었다. 해당 시스템으로는 Ariadne, HERMES, InfoSleuth, Information Manifold, 그리고 Whirl 등이 있다[10-12]. 특히 MCC의 InfoSleuth 프로젝트는 Carnot의 기능을 동적으로 변화하는 환경으로 확장시켰다. InfoSleuth는 고수준 질의어인 KQML을 통하여 통신하는 협조 에이전트 네트워크로 구성된다[13]. 이러한 시스템들은 유연성과 개방성을 지원하기 위하여 설계되었으나 사용자가 정보를 통합하기 위하여 필요한 도메인의 지식을 알고

있어야 하는 부담이 존재한다. 최근에는 온라인 사전과 시소러스(thesaurus)의 발전으로 사용자가 언어 이론을 토대로 편리하게 정보를 통합할 수 있도록 지원한다. 대표적인 시스템으로는 Summary Schemas Model이 있으며 전역데이터 구조를 사용한다. 이것은 정보를 접근할때 시스템 지정적인 용어가 아닌 사용자 자신의 용어 사용을 허용한다. 그러나 이러한 접근 방법은 질의를 단일의 논리적 인덱스에 한정하기 때문에 규모가 큰 멀티 데이터베이스 시스템에서는 성능의 한계를 가진다. 이렇게 다양하게 접근된 정보 통합 방법들의 궁극적인 목적은 의미 이질성을 해결하는 것이고 본 논문에서 다루고자 하는 MDR과 온톨로지를 결합한 정보 통합 시스템에 대한 기술요소를 다음과 같이 정리할 수 있다.

2.1 스키마 및 데이터 이질성 분류

멀티 데이터베이스 시스템에서 발생할 수 있는 이질성을 분류하는데 있어서 플랫폼 이질성과 의미 이질성으로 크게 나눌 수 있다. 플랫폼 이질성은 많은 연구를 통해서 해결이 되어왔으나 의미 이질성은 아직도 해결해야 할 과제로 남아 있다. 의미 이질성은 다시 스키마 이질성과 데이터 이질성으로 분류할 수 있다. 스키마 이질성은 주로 동일한 정보에 대해 다른 구조를 사용하고 동일한 구조에 대해서 서로 다른 개체명이나 속성명을 사용함으로써 발생한다. 데이터 이질성은 스키마 이질성이 해결된 상태에서 동일한 개념의 데이터

<표 1> 의미 이질성의 분류

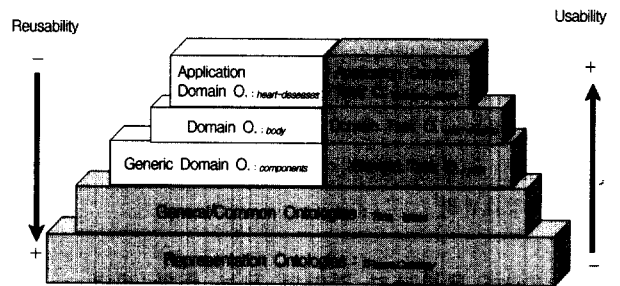
이질성 종류	분 류	
	충돌 종류	정 의
스키마	개체간 구조 충돌	서로 다른 요소 데이터베이스가 동일한 정보를 표현하는데 있어 다른 개수의 개체를 사용할때 발생
	개체와 속성간 구조충돌	어떤 데이터베이스의 속성이 다른 데이터베이스에서는 개체로 표현될때 발생
	개체와 값간 구조 충돌	어떤 데이터베이스의 속성값이 다른 요소 데이터베이스에서 개체와 의미적으로 연관될때 발생
	속성간 구조 충돌	서로 다른 요소 데이터베이스가 동일한 정보를 표현하는데 있어 서로 다른 개수의 속성을 사용할때 발생
	속성과 값간 구조충돌	어떤 데이터베이스의 속성값이 다른 데이터베이스에서 속성과 의미적으로 연관될때 발생
	개체간 이름 충돌	서로 다른 데이터베이스에서 동일한 정보를 표현하는 개체에 대해 다른 이름을 사용할 때 발생
	속성간 이름 충돌	동일한 정보를 표현하는 속성에 대해 다른 이름을 사용할때 발생
데이터	잘못된 데이터	어떤 데이터베이스의 관리적인 측면에서 날짜나 무결성 제약 등의 문제가 생길때 발생
	동일데이터에 대한 다른 표현	요소 데이터베이스간의 같은 데이터에 대해 표현, 단위 및 정밀도의 차이 때문에 발생

에 대한 표현의 불일치로 발생한다. 세부적인 분류는 <표 1>과 같다[4, 14].

본 논문에서 스키마 이질성에 대한 해결은 온톨로지를 이용하여 자동적으로 사상하며 데이터 이질성에 대한 해결은 메타데이터 저장소(Metadata Repository)와 지식 베이스를 구축하여 반자동적으로 해결한다.

2.2 온톨로지(Ontology)

온톨로지는 지식베이스에 표현되어 있는 지식 이면의 개념을 명백히 기술하기 위해 계층적으로 구조화된 용어의 집합이다[15]. 온톨로지는 주제영역의 단어로 구성된 기본용어와 용어간 관계를 정의한다. 용어들과 관계를 조합하기 위한 규칙뿐만 아니라 단어의 확장을 정의한다. 구성요소는 다양한 형태의 용어로 쓰여지지만 범용적으로 다음과 같이 개념(Concept), 관계(Relation), 함수(Function), 공리(Axiom) 등의 요소로 나타낼 수 있다. 적용범위는 시스템간의 상호운용성, 사람과 시스템간의 통신 및 명세나 재사용 요소 신뢰성 측면의 시스템 공학에 적용하며 실제 적용 시스템은 TOP, WordNet, Ontolingua, Cyc, Frame Ontology, PAN-GLOSS, MikroKosmos, SENSUS, EngMath, PhysSys, TOVE, CHEMICALS 등이 있다[1, 7]. 이러한 온톨로지는 내용 중심 온톨로지와 개념 중심 온톨로지로 구분할 수 있다. (그림 1)에서 처럼 이러한 구분은 재사용성과 적용성 측면에서 반대의 경향을 보인다. 온톨로지의 재사용성이 높아질수록 적용성 측면은 낮아지고 적용성이 높아질수록 재사용성은 낮아진다.



(그림 1) 온톨로지 타입 분류 및 적용범위

본 논문에서는 사용자 질의를 기본 단위로 분해하여 개념을 추출하는 용도와 MDR의 데이터요소간 관련성을 검색하여 정보를 통합하기 위한 의미망으로서 온톨로지를 적용한다.

2.3 MDR(Metadata Registry)

MDR에 대해서는 ISO 11179에서 정의하고 있다. ISO 11179는 메타데이터 즉, 데이터요소의 명세와 표준화를 위하여 데이터의 요청과 등록을 유용화하고, 데이터의 접근과 사용을 촉진하기 위하여 데이터를 이해할 수 있고, 공유할

수 있도록 만들기 위한 표준과 등록에 대한 내용을 설명하고 있다[2].

MDR은 메타데이터의 등록과 인증을 통하여 표준화 된 메타데이터를 유지 관리하며, 메타데이터의 명세와 의미 공유를 목적으로 한다. 따라서, 이것은 데이터 공유를 위한 기본 틀로서 공유를 위한 기본요소로 데이터 요소를 사용하고, 이러한 데이터 요소가 그 내포적인 의미를 충분히 나타내어 줄 수 있는 구조 내에 등록하여 각 응용에 활용할 수 있다. MDR은 관리부 (stewardship region), 명명과 식별부분(naming & identification region), 분류부분(classification region), 데이터 요소 관리부분(data element administration), 데이터 요소 개념 관리부분(data element concept administration), 그리고 개념 또는 값 영역 관리부분(concept or value domain administration)으로 구성된다.

MDR을 적용한 시스템으로는 미국의 EPA(Environmental Protection Agency)에서 이해가 쉽고 신뢰성이 높은 환경관련자료를 제공하고 여러 관련 기관들의 표준화 작업을 돕기 위해 EDR(Environmental Data Registry)을 구축하여 기관들간 데이터 공유를 가능하게 했다. 미 교통부(the Department of Transportation)에서는 IEEE P1489를 적용한 ITS(Intelligent Transportation System)를 구축하여 ITS의 각 시스템 사이의 상호 정보교환을 가능하게 했다. 미 건강국(HA : Health Affair)에서 USHIK(U.S Health Information Knowledgebase)를 구축하여 국민 보건에 관련된 데이터의 공유를 보장한다[16].

본 논문에서 적용되는 MDR은 다음과 같은 역할을 수행한다.

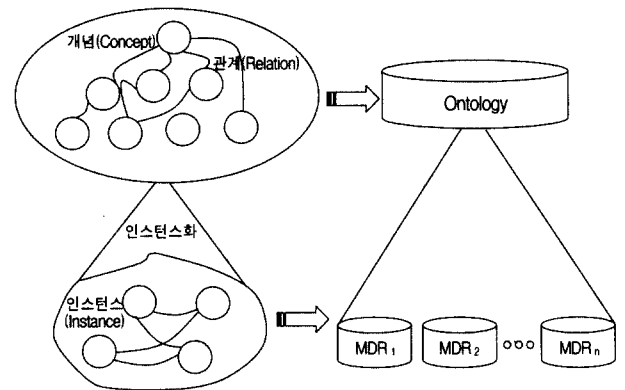
- 하위의 데이터베이스를 통합하기 위한 표준 정보제공
- 데이터요소와 하위 데이터베이스 스키마 사이의 사상 정보 유지
- 사용자 질의에 사상되어 있는 개별 데이터베이스 스키마 정보관리

3. MDR과 온톨로지 결합의 정보 통합 개념

실세계의 사물을 컴퓨터상의 활용 가능한 자원으로 만들기 위해서 추상화의 단계를 필요로 하게 된다. (그림 2)와 같이 실세계의 객체들은 개념으로 추상화 될 수 있다. 클래스와 인스턴스의 개념과 유사하다. 또한, 객체들간의 상호작용은 몇 가지 유형의 관계로 사상될 수 있다. 이렇게 객체의 추상화된 개념과 관계의 정보를 유지하는 것이 온톨로지가 된다. 온톨로지는 마치 사전처럼 쉽게 변하지 않고 사용자가 원하는 정보를 질의할때 질의에서 유사한 개념을 추출하기 위한 방법이 된다. 개념은 하나로 추상화 되지만 해당 개념을 표현하는 객체는 다양한 형태일 수가 있다.

MDR은 이런 다양한 객체들의 표현을 표준화된 형태로 유

지한다. 표준화되어 있다는 의미는 한 도메인 영역 내에 다양하게 사용되는 데이터들의 내용을 대표적인 표현으로 정의한다는 것이다. (그림 2)에서, MDR간의 표준화된 데이터 요소의 표현이 단일의 온톨로지 개념으로 사상되는 경우를 볼 수 있다. 위 구조에서 고려해야 할 문제는 MDR간 데이터 요소의 이질성이 발생할 경우이다. 본 논문에서는 주로 데이터 이질성의 유형에 따라 에이전트가 지식베이스를 이용하여 이질성을 해결하게 된다. 이렇게 구성된 온톨로지와 지식 베이스는 사용자의 범용적인 질의를 재구성하는데 사용된다. 온톨로지에서 개념검색을 통한 1차 개념 질의 분할이 수행되고 MDR의 데이터 요소들간의 사상정보와 지식 베이스의 검색을 통해서 2차 메타데이터 요소 질의분할이 수행된다.



(그림 2) MDR과 온톨로지 기반의 정보 통합 개념도

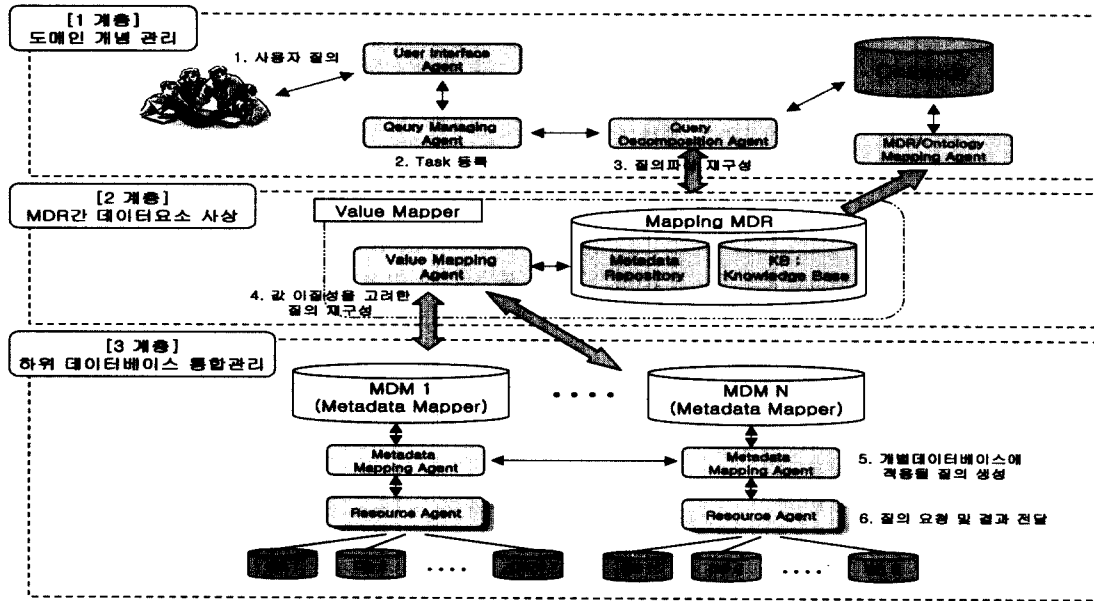
4. LI2S 아키텍처

MDR과 온톨로지 기반의 정보 통합 개념을 구체화한 LI2S 아키텍처에 대해 소개한다. 그리고 계층별 에이전트와 핵심 기능에 대해서 자세히 기술한다.

4.1 시스템 개요

3장 “MDR과 온톨로지 결합의 정보 통합 개념”에서 언급한 내용을 적용하면, 온톨로지를 이용해 의미망을 형성하여 정보 통합을 위한 기반을 마련하고 MDR이 가진 정보공유와 표준성을 이용한 정보 통합 시스템을 구축할 수 있다. 핵심 요소는 크게 온톨로지, VM(Value Mapper), MDM(Metadata Mapper)으로 나눌 수 있다. 각 핵심요소의 역할을 중심으로 계층화하여 독립성을 유지하면 시스템 변경에 대한 유연성을 확보할 수 있다. 이러한 계층별 독립성을 제공하기 위해 지능성을 가진 에이전트가 요구된다[17].

본 논문에서 제안하는 MDR과 온톨로지를 결합한 3계층 정보 통합 아키텍처는 (그림 3)과 같이 핵심요소를 중심으로 3계층으로 구성되어 있으며 아래와 같이 각 구성요소와 계층 내에서 작동하는 에이전트가 존재한다.



(그림 3) L2S 아키텍처 및 통합과정

4.2 제 1계층 - 도메인 개념관리

도메인의 개념을 관리하며 MDR과 온톨로지의 지식 베이스를 기반으로 지식 베이스에 갱신되는 내용을 통합 관리자에 제공하여 사상을 위한 규칙을 수정하는 반자동적인 방식을 사용한다.

4.2.1 UIA(User Interface Agent)

UIA는 사용자의 질의를 파싱하며 온톨로지 상의 개념과 관계를 검색하여 사상되어 있는 MDR의 데이터요소 개념을 추출하여 질의를 요청하고 처리된 질의 결과를 사용자에게 보여주는 역할을 한다.

4.2.2 QMA(Query Managing Agent)

QMA는 사용자가 요청한 다수의 질의에 대해 쓰레드를 이용한 작업의 생성/추적/종료 등의 관리를 담당한다.

4.2.3 QDA(Query Decomposition Agent)

QDA는 온톨로지에 Mapping되어 있는 MDR 데이터요소의 정보를 이용하여 사용자의 질의를 부분질의로 분해한다.

4.2.4 MOMA(MDR/Ontology Mapping Agent)

MOMA는 새로운 데이터 소스의 통합에 의해 MDR의 데이터 요소 개념을 추가해야할 경우, MDR과 온톨로지간 새로운 사상작업을 수행하는 에이전트이다. 이러한 사상 작업은 다음과 같은 MDR과 온톨로지 사상규칙을 정의하여 사용한다.

- 온톨로지에 등록된 개념이 MDR에 없는 경우
온톨로지를 구성하는 측면에서, MDR에 개념이 존재하는 것에 대한 검색은 불필요하다. 온톨로지는 도메인 관리자에 의해 새로운 개념이 추가되고 개념간 관계가 재 설정될 수 있다.

- MDR에 존재하는 개념이 온톨로지에 없는 경우
MDR에 새로운 데이터 요소가 추가되는 경우가 해당되며 온톨로지에 후보 개념으로 등록된다.
- 온톨로지에 개념이 존재하는 경우
온톨로지를 검색하여 MDR에 등록된 데이터 요소와 온톨로지 상의 개념과 관계를 확인하여 사상 관계를 설정한다.

4.2.5 온톨로지

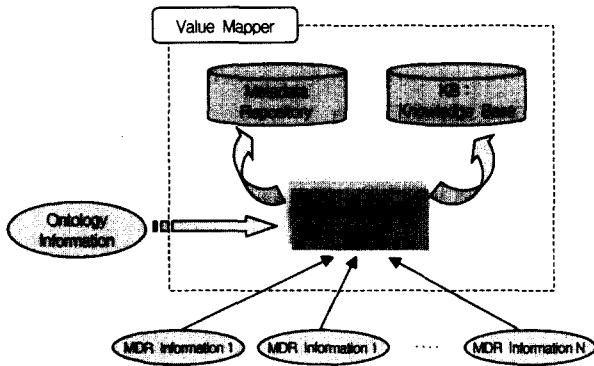
온톨로지는 적용하고자 하는 도메인에 필요한 개념들을 추출하고 개념들간의 관계를 설정하여 유지하는 역할을 한다. 사용자의 질의에서 추출된 요소들과 검색을 할 수 있는 일종의 의미망(Semantic network)이다.

- 구성 요소
 - 개념(Concept) : 특정 도메인에서 사용되는 일반화된 용어
 - 서술(Description) : 개념의 정의를 표준화된 용어로 기술하는 것
 - 관계(Relation) : 개념들간의 관련 유형을 정의하여 연결시키는 것
- 관계 유형
 - 동등(E-R) : 개념간의 의미가 동일한 경우
 - 상속(IS-A) : Class와 Instance와 같은 경우
 - 부분(HAS-A) : A has a B로 A가 B를 부분으로 가진 경우

4.3 제 2계층 - MDR간 데이터요소 사상

온톨로지의 개념 및 관계 정보를 이용하여 MDR간 데이터 요소 사상을 담당하며 사상 MDR과 VMA(Value Map-

ping Agent)로 구성된 VM(Value Mapper)를 구성한다. VM은 (그림 4)에서 처럼, 메타데이터 저장소, 지식베이스 및 VMA (Value Mapping Agent)로 구성된다. VM은 1계층의 온톨로지 정보와 3계층의 MDR 정보를 참조하여 메타데이터 저장소에 데이터요소간 사상 작업을 자동으로 수행하며 지식베이스를 구축하기 위한 인터페이스를 제공한다.



(그림 4) VM(Value Mapper) 구조도

(그림 4)에서 보여지는 구성요소의 설명, 데이터요소 값 간 사상유형과 사상규칙에 대한 설명은 다음과 같다.

4.3.1 메타데이터 저장소(Metadata Repository)

개별 MDR의 데이터 요소(Data Element)와 속성(Property)들에 대한 정보를 저장하고 온톨로지서 동일 개념으로 관계를 설정한 데이터 요소들간의 사상정보를 유지한다.

4.3.2 지식 베이스(KB : Knowledge Base)

데이터요소 표현에서 발생하는 데이터 이질성 해결을 위한 변환 데이터와 정의 규칙을 구성한다.

4.3.3 VMA(Value Mapping Agent)

VMA는 두 가지 역할을 한다. 하나는, 메타데이터 저장소에 데이터 요소간 사상정보를 저장하고 지식베이스에 MDR 간 데이터 값에 대한 사상을 담당한다. 다른 하나는, MDM (Metadata Mapper)에서 처리 될 개별 MDR 질의를 생성한다. 질의 생성시 메타데이터 저장소와 지식 베이스를 이용하여 MDR간 동일한 개념의 데이터요소를 나타내는 각 데이터요소간 Value 이질성을 해결한다[5].

4.3.4 데이터 요소 값간 사상유형과 사상규칙

데이터 요소간 값간 사상은 3가지 사상 유형과 다양한 사상 규칙이 존재한다. 사상유형은 크게 분류, 집합, 일반의 3가지 계층적 사상유형과 치환, 전환의 2가지 비 계층적 사상유형으로 분류할 수 있다. 본 논문에서는 치환과 전환의 비계층적 사상유형에 대한 문제점을 해결하기 위해 초점을 맞추었다. <표 2>에 사상규칙에 대한 예가 제시되었다[18]. 변환 데이터와 정의 규칙들은 다음과 같은 데이터 유형의 표현에 대한 것이다.

<표 2> 비계층적 사상유형 분류 및 예

중 분류	소 분류	분류 예
비계층적 사상유형	정적전환유형	M/cm, cm/M
	동적전환유형	W/\$, \$/W
	치환 유형	조선시대, 1392, 1910

즉, 정적전환유형은 거의 변하지 않는 전환규칙이고 원화와 달러와 같이 전환규칙이 변할 수 있는 경우를 동적전환 유형이라고 한다. 치환유형은 예와 같이 조선시대는 1392년부터 1910년까지라는 전환규칙이 존재하지 않는 대치되어야 할 유형에 대해 정의한다. 이와 같이 일반적으로 사용되어 지는 데이터 표현은 모두 처리가능하다.

4.4 제 3계층 - 하위 데이터베이스 관리

실제 연동되는 하위 데이터베이스들에 대한 통합관리를 담당한다.

4.4.1 MDMA(Metadata Mapping Agent)

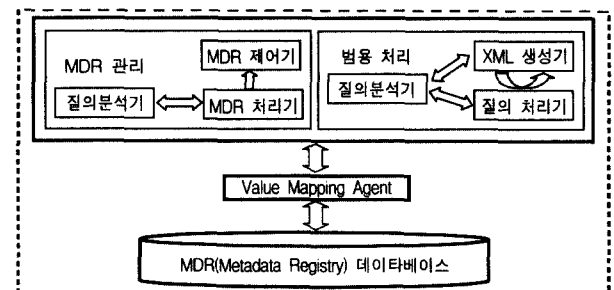
MDMA는 MDM과 RA 사이에 존재하여 메타데이터와 각 개별 스키마 정보를 사상시키며, 변경되는 정보를 VMA에 제공하는 역할을 한다.

4.4.2 RA(Resource Agent)

RA는 각 데이터베이스별로 존재하는 에이전트로 질의 문의 형태를 실제 데이터베이스에 맞도록 변경 질의하며 그 결과값을 MDM에 전달하는 역할을 한다.

4.4.3 MDM(Metadata Mapper)

각 MDR내에는 표준화되어 있는 데이터요소가 존재하고 이 데이터요소에 사상되어 있는 하위 데이터베이스의 스키마 정보를 식별할 필요가 있다. (그림 5)와 같이 MDM(Metadata Mapper)은 크게 사용자 질의에 대한 처리를 위한 범용처리 부분과 RA(Resource Agent)로부터 보고되는 스키마 정보를 이용하여 반자동적으로 MDR의 데이터요소와 사상시키는 MDR 관리 부분으로 구성되며 그 기능은 다음과 같다.



(그림 5) MDM(Metadata Mapper) 아키텍처

4.4.4 MDR 관리부

질의분석을 통해 MDR 관리자를 식별하여 MDR의 데이

터 요소를 추가 삭제하는 기능을 부여하는 MDR 제어기와 그 기능을 처리하는 MDR 처리기로 구성된다.

4.4.5 범용 처리부

질의가 XML 생성 요청과 일반 질의인지를 분석하는 질의분석기, 웹 환경에서 데이터 공유를 위한 표준으로 많은 영역에서 자리잡고 XML 스키마를 생성하여 제공할 수 있는 XML 생성기 그리고 각 Resource Agent로 보내질 질의를 생성 처리하는 질의 처리기로 구성된다.

5. LI2S 설계 및 구현

본 장에서는 4장에서 언급된 LI2S 아키텍처에 대해 요구 사항을 정의하고 적용사례 분석을 통한 설계를 수행하였다. 오라클과 인터베이스 등의 데이터베이스를 이용한 실제 적용 사례를 통해 시스템을 구현한다.

5.1 설계시 고려사항

LI2S 시스템 설계시 고려한 행위자는 데이터요소 관리자, 하위 데이터베이스 관리자 및 사용자이다. 데이터요소 관리자는 온톨로지의 개념 및 관계를 정의하는 일부 데이터요소간 이질성 해결을 위한 지식베이스 구성에 책임을 지는 사람이다. 사용자는 제공되는 사용자 인터페이스(User Interface) 상에서 자신이 알고 있는 개념을 이용해 질의를 하고 통합된 정보를 이용하는 사람이다. 하위 데이터베이스 관리자는 RA(Resource Agent)에 자신이 관리하는 데이터베이스 스키마 정보와 MDR의 데이터 요소간 사상을 담당

한다. 설계시 MDR의 데이터요소 등록과 데이터 요소간 이질성해결 부분과 온톨로지를 구성하여 MDR의 데이터 요소간 사상에 초점을 맞추었으며 MDR의 데이터 요소와 하위의 데이터베이스 스키마 정보는 사상되어있다는 가정을 하였다. 이러한 가정을 통해 고려한 프로세스는 다음과 같다.

- 온톨로지 구성 및 데이터요소 등록에 따른 통합 MDR 재구성
- MDR의 데이터요소 등록에 따른 데이터 이질성 해결을 위한 지식 베이스 재구성
- 사용자 질의에 대한 계층별 질의 분해 및 결과 통합

5.2 적용 사례

한국과학기술정보연구원(KISTI)에는 국내에서 과학기술 정보와 관련된 다양한 분야의 데이터베이스를 보유하고 있

<표 3> 적용 도메인의 데이터베이스

데이터베이스 이름	데이터베이스 설명
DIGS	• 정기 간행물 데이터베이스 • 외부에서 도입된 데이터베이스
DIMD	• 석박사 학위 논문
BIST	• 과학기술정보 분야 • 과학기술분야의 주요 잡지에 대한 서지정보
INFO	• 문헌정보학 분야 • 정보관리 분야의 주요 잡지에 대한 서지정보
THESIS	• 석박사 학위 논문
KRIST	• 연구 개발 보고서
SOCIETY	• 학회지 및 학술대회논문집

<표 4> 추출 개념과 데이터베이스 스키마 사상표(일부)

Concept	DICS		DIMD		BIST		INFO		THESIS		KRIST		SOCIETY	
	Field	Type	Field	Type	Field	Type	Field	Type	Field	Type	Field	Type	Field	Type
문헌번호	F_DN	문헌번호	F_DN	문헌번호	F_DN	문헌번호	F_DN	문헌번호	AN	제어번호	AN	제어번호	AN	제어번호
발행연도	F_PY	발행연도	F_PY	발행연도	F_PY	발행연도	F_PY	발행연도	PY	학위연도	PB	발행일자	PY	발행연도
발행국	F_CY	발행국	F_CY	발행국	F_CY	발행국	F_CY	발행국						
사용언어	F_LA	사용언어	F_LA	사용언어	F_LA	사용언어	F_LA	사용언어	LA	사용언어	LA	사용언어	LA	언어
학위구분		학위구분	F_DG	학위구분					DK	학위구분				
한글제목	F_TI	제목	F_TI	제목	F_TI	제목	F_TI	제목	TI	논문제목	TIK	연구과제명(한글)	TI	원문제목
저자명	F_AU	저자	F_AU	저자	F_AU	저자	F_AU	저자	PNH	개인저자	PNH	주관연구 기관연구 책임자	PNH	저자
분류	F_CC	주제분류	F_CC	주제분류	F_CC	분류	F_CC	분류						
색인어	F_KW	색인어	F_KW	색인어	F_KW	색인어	F_KW	색인어	ID	저자키워드	ID	저자키워드	KW	키워드
한글초록			F_AB	초록	F_AB	초록	F_AB	초록	ABK	한글초록	ABK	요약문(한글)	AB	초록
페이지					F_SP	페이지	F_SP	페이지	PH	페이지	PH	페이지수	PG	페이지
이미지연수									PHI	이미지연수	TPG	이미지수		
자료형태					F_DT	자료형태	F_DT	자료형태			DT	자료유형	DT	자료유형
청구기호					F_CL	청구기호	F_CL	청구기호	CN	청구기호				
소장기관					F_CP	소장기관					LO	소장기관	LO	소장위치
서지정보	F_SO	서지사항	F_SO	서지사항										
참고문헌수											RF	참고문헌수	RF	참고문헌수

다. 그러나 각 분야만을 고려하여 설계되고, 유사한 분야일 경우에도 서로 다른 표현의 데이터베이스 스키마를 정의하여 사용함으로써 스키마 이질성과 데이터 이질성이 발생한다. 또한 비정규화된 설계로 인해 중복성 문제와 일관성 있는 관리가 어렵다. 즉, 이질적인 데이터베이스들의 통합에서 일관성을 유지하기 어렵고 상호 정보교환이 어렵다는 문제점을 가지고 있다. <표 3>은 서지정보 도메인에서 분석 대상으로 선정된 테이블이다. 물리적으로는 오라클 데이터베이스와 크리스탈 데이터베이스로 구성되어 있으며 서울과 대전 두 지역에 분산되어 관리되고 있다[20].

이러한 문제점을 해결하기 위해 서지정보 분야의 데이터베이스에 대한 스키마 및 데이터 분석을 통해 <표 4>와 같은 표준화된 59개의 메타데이터 요소를 추출하였다.

이렇게 추출된 데이터요소는 <표 5>와 같이 데이터요소

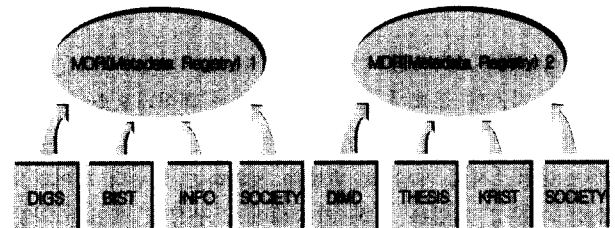
<표 5> 문헌번호(Record_Reference_Number)에 대한 속성 정의서

데이터요소의 메타데이터 속성정의서		
식별형 속 성	(1) 명칭(Name)	Record Reference Number
	(2) 식별자(Identifier)	PR11
	(3) 버전(Version)	1.0-2002
	(4) 등록기관(Registration Authority)	KISTI
	(5) 동의명(Synonymous Name)	문헌번호
	(6) 문맥(Context)	번호
정의형 속 성	(7) 정의(Definition)	문헌정보의 식별자로서 KISTI에서 부여하는 유일한 값을 가진다.
관계형 속 성	(8) 분류체계(Classification Scheme)	서지정보-식별
	(9) 핵심어(Keyword)	문헌번호
	(10) 개념영역(Conceptual Domain)	도서
	(11) 개념영역 정의 (Conceptual Domain Definition)	글씨, 그림, 책 등의 총칭
	(12) 개념영역명(Conceptual Domain Name)	도서
	(13) 객체부류(Object Class)	식별클래스
	표현형 속 성	(14) 표현범주(Representation Category)
(15) 표현양식(Form of Representation)		본 문
(16) 값 형태(Value Format)		문자열
(17) 값 최대길이(Maximum Length)		30
(18) 값 최소길이(Minimum Length)		1
(19) 표현의 배치(Layout of Representation)		없 음
(20) 허용범위(Permit Value)		표기하지 않음
관리형 속 성		(21) 책임조직(Responsible Organization)
	(22) 등록상태(Registration Status)	등록
	(23) 발의조직(Submitting Organization)	KISTI
	(24) 변경일자(Change Date)	20021001
	(25) 변경사항 기술문 (Change Description Text)	신규등록
	(26) 생성자명(Create by User Name)	정동원
	(27) 생성일시(Create Date)	20021001
	(28) 책임자명 (Data Steward Organization Name)	백두권

자체에 대한 속성 정의를 통해 다음과 같이 표준화된 형태로 표현될 수 있다. 즉, 문헌번호라는 메타데이터요소는 명칭이 'Record Reference Number' 또는 '문헌번호'로 사용되어지고 값 형태는 문자열로서 최대 30자 최소 1자의 범위로 표현될 수 있다[20].

이러한 문헌번호를 검색하여 참조하고자 하는 사용자는 해당 속성정의에 대한 규격을 준수하여야 한다. 이렇게 추출된 메타데이터요소는 기존 분산되어 처리되고 있는 데이터베이스의 물리적 구조와 지역 분산 특징을 고려해 서지정보에 대한 메타데이터를 (그림 6)과 같이 2개의 MDR을 이용해 구성하였다.

(그림 6)에서, MDR 1은 서지정보에 있어서 정기 간행물, 과학분야에 대한 간행물과 문헌정보에 대한 메타데이터를 구성한다. MDR 2는 석·박사 학위 논문과 연구개발 보고서에 대한 서지정보를 구성한다.



(그림 6) MDR 구성 개념도

이렇게 구성된 MDR1과 MDR2에 대한 메타데이터요소에 대한 정보는 <표 6>, <표 7>과 <표 8>에서처럼 LIS의 2계층의 Metadata 저장소를 구성하는 MDR과 지식 베이스를 구성하는 테이블로 구성된다.

<표 6> MDR 테이블 구조(일부)

Attribute	DataType	Constraint	Description
tmp_id	VARCHAR2(15)	Primary key	식별자
concept_name	VARCHAR2(30)	NOT NULL	요소이름
registration	VARCHAR2(30)	NOT NULL	등록기관
synonymous	VARCHAR2(100)		요소와 같은 의미의 다른 요소명
...
object_class	VARCHAR2(30)	NOT NULL	객체 분류
category	VARCHAR2(30)	NOT NULL	표현 범주
formation	VARCHAR2(30)	NOT NULL	표현 양식
data_type	VARCHAR2(30)	NOT NULL	데이터 요소 데이터 유형
max	NUMBER(4)	NOT NULL	데이터 요소 최대 크기
min	NUMBER(4)	NOT NULL	데이터 요소 최소 크기
layout	VARCHAR2(30)		표현의 배치
permissible_data	VARCHAR2(30)	NOT NULL	허용 가능한 데이터 요소 값
create_user	VARCHAR2(30)	NOT NULL	생성자명
create_date	DATE	NOT NULL	생성일시
steward_name	VARCHAR2(30)	NOT NULL	책임자명

<표 7> 온톨로지 개념과 데이터요소 사상 테이블 구조

Attribute	DataType	Constraint	Description
mapping_id	VARCHAR2(20)	Primary key	사상을 위한 식별자
concept_name	CHAR(10)	.	온톨로지의 개념명
DE_name	VARCHAR2(30)	.	MDR의 데이터요소명
relation_type	VARCHAR2(30)	.	개념과 데이터요소와의 관계

<표 8> MDR간 데이터요소 데이터 이질성에 대한 지식베이스 테이블 구조

Attribute	DataType	Constraint	Description
mapping_id	VARCHAR(10)	Primary key	사상을 위한 식별자
data_element	VARCHAR(30)	.	변환주체 데이터요소
trans_unit	VARCHAR(30)	.	변환 단위
trans_type	CHAR(2)	.	변환 유형
operation	CHAR(3)	.	연산자
trans_value	CHAR(10)	.	변환 값

구축한 MDR의 데이터요소 중 지식베이스로 구성할 수 있는 데이터요소와 실제 표현은 다음과 같다.

- ① 가격: 기준은 원화를 기본으로 하나 외국 서적의 경우 예들들어, \$나 ¥로 표현될 수 있으며 동적인 변환 유형이다.
- ② 발행국: 대한민국에 대한 표기를 'KOR'과 'KOREA' 두가지를 혼용하여 사용하고 있으며 치환유형으로 분류한다.

<표 9> 구성 지식베이스의 내부 표현

mapping_id	data_element	trans_unit	trans_type	operation	trans_value
kb0001	cost	\$	VT	×	1300
kb0002	publish_country	KOR	TT		KOREA
kb0003	cost	¥	VT	×	1000
kb0004	publish_country	ENG	TT		ENGLISH

MDR간 데이터요소 데이터 이질성을 해결하기 위해 구성한 지식베이스를 검색하여 관련된 데이터요소의 규격변환에 대한 알고리즘은 <표 10>과 같다.

<표 10> 지식베이스를 이용한 데이터요소 규격변환 알고리즘

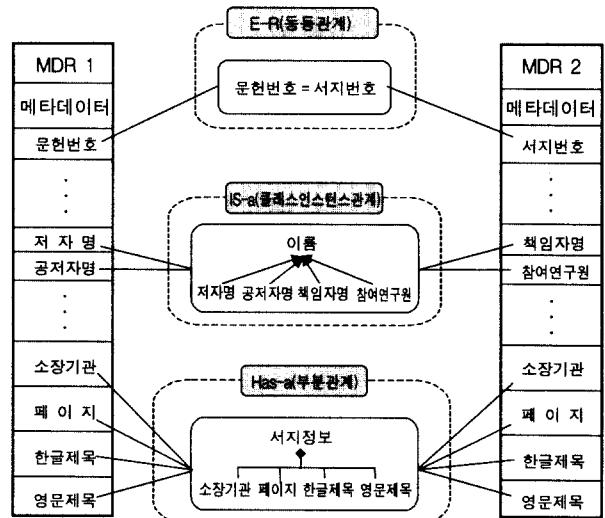
```

WHILE (KB_identifierr < > Null)
BEGIN
  GET MDR Data Element /* 검색 데이터 요소에 대한 정보를 획득한다. */
  FIND trans_type /* 변환 유형에 대해 결정한다. */
  IF ((trans_type = VT) or (trans_type = ST))
  /* 변환 유형이 동적변환 or 정적변환인가? */
  BEGIN
    COMPARE selected unit to trans_unit
    /* 선택되어진 규격단위와 검색되어진 규격단위를 비교한다. */
    FIND operation and value /* 검색되어진 규격단위의 연산
    
```

```

자과 값을 찾는다. */
CHANGE data element value /* 데이터 요소 변수의 값을 변경한다. */
SET Query division /* 질의를 분할한다. */
END
ELSE IF( trans_type = TT) /* 변환 유형이 치환유형인가? */
BEGIN
  COMPARE data element vale to trans_unit
  /* 데이터 요소 변수 값과 규격단위를 비교한다. */
  IF( data_element_value = trans_unit) /* 일치되는가? */
  BEGIN
    SET Query division /* 질의를 분할한다. */
  END
END
END WHILE
    
```

LI2S 1계층의 온톨로지를 구성하기 위하여 (그림 7)과 같이 MDR1과 MDR2의 메타데이터 요소의 동의명만을 추출하여 나열하였다. 나열된 데이터 요소들은 온톨로지의 개념으로 추출되고 관계 유형에서 정의한 3가지 형태로 온톨로지의 개념과 연관될 수 있다.



(그림 7) MDR간 개념 사상방법

(그림 7)은 이러한 MDR간의 메타데이터 요소에 대한 사상 관계를 도식적으로 보여준다. 이렇게 추출된 개념과 관계에 대한 일부는 다음과 같다.

- ① E-R(동등관계)
 - 문헌번호와 서지번호는 동등한 관계이다.
 - 저자명과 책임자명은 동등한 관계이다.
 - 발행국과 발행처는 동등한 관계이다.
 - 가격과 서지가격은 동등한 관계이다.
- ② IS-A(상속관계)
 - 이름과 저자명, 공저자명, 책임자명, 참여연구원은 상속관계이다.
 - 제목과 한글제목, 영문제목은 상속관계이다.

③ HAS-A(부분관계)

- 서지정보에 대한 구성정보는 소장기관, 페이지, 한글제목, 영문제목이다.

<표 11> 온톨로지 테이블 내부표현(일부)

Onto_id	Concept_M	Relation	Concept_S
Ont00001	문헌번호 (Record Reference Number)	E-R	서지번호 (Bibliography Number)
Ont00002	서지번호 (Bibliography Number)	E-R	문헌번호 (Record Reference Number)
Ont00003	저자명 (Author Name)	E-R	책임자명 (Person Name)
Ont00004	발행국 (Publication Country)	E-R	발행처 (Publication Department)
Ont00005	이름(Name)	IS-A	저자명(Author Name)
Ont00006	이름(Name)	IS-A	공저자명(Coauthor Name)
Ont00007	이름(Name)	IS-A	책임자명(Person Name)
Ont00008	이름(Name)	IS-A	참여연구원(Researcher Name)
Ont00009	서지정보 (Bibliography Information)	HAS-A	소장기관 (Possession Organization)
Ont00010	서지정보 (Bibliography Information)	HAS-A	페이지(Page)
Ont00011	서지정보 (Bibliography Information)	HAS-A	한글제목(Title Korean)
Ont00012	서지정보 (Bibliography Information)	HAS-A	영문제목(Title English)
Ont00013	가격(cost)	E-R	서지가격(pricing)
Ont00014	제목(Title)	IS-A	한글제목(Title Korean)
Ont00015	제목(Title)	IS-A	영문제목(Title English)

이렇게 구성된 온톨로지는 사용자의 의미질의에 대해 분석을 위한 의미망으로 사용되며 LI2S 2계층의 Value Mapper에 메타데이터 재구성에 대한 정보를 제공하게된다. <표 12>는 온톨로지 정보를 이용하여 메타데이터를 재구성하기 위한 알고리즘이다.

<표 12> 메타데이터 저장소에서 데이터 요소간 사상 알고리즘

```

WHILE (Monitor < > Fail)
BEGIN
    GET MDR Data Element /* 데이터 요소에 대한 정보를 획득한다. */
    FIND Concept in the Ontology /* 온톨로지상에서 대응되는 개념을 찾는다. */
    IF (Data Element = Concept)
    BEGIN
        SEARCH Concept with Relation of the Concept /* 개념과 관계를 가진 개념을 검색한다. */
        FIND Corresponding Data Element Concept with Relation /* 관계를 가진 개념에 대응되는 데이터 요소를 찾는다 */
        IF (Relation = E-R Relation) /* E-R 관계이면 해당 개념을 메타데이터 저장소에 저장한다 */
    
```

```

BEGIN
    SET Data Elements to Metadata Repository with E-R Relation
END ELSE IF(Relation = Is-a Relation)
/* Is-a 관계이면 해당 개념을 메타데이터 저장소에 저장한다 */
BEGIN
    SET Data Elements to Metadata Repository with Is-a Relation
END ELSE /* Has-a 관계이면 해당 개념을 메타데이터 저장소에 저장한다 */
BEGIN
    SET Data Elements to Metadata Repository with Has-a Relation
END
END ELSE
BEGIN
    NOTIFY No Concept to Ontology Administrator
    /* 온톨로지 관리자에게 개념이 존재하지 않음을 통보한다.
END
END WHILE
    
```

5.3 정보 통합 과정

다양한 정보의 통합을 제공하는 LI2S 시스템을 이용하여 사용자는 스키마 정보를 알지 못하더라도 자신이 알고 있는 개념을 이용하여 질의를 할 수 있다. 의미질의를 위해 FROM 절이 없는 것을 제외하고는 SQL과 유사한 SemQL(Semantic Query Language)을 도입한다[19]. SemQL의 기본형태는 두 개의 절로 구성되고 다음과 같은 형태를 가진다.

```
SELECT <concept list> WHERE <condition>
```

<concept list>는 질의에 의해 검색되어야 하는 값의 개념에 대한 리스트이다. <condition>은 질의에 의해 검색되어야 하는 튜플을 식별하기 위한 조건식이다.

통합된 정보를 검색하고자 하는 사용자들은 SemQL로 자신들이 필요로 하는 질의를 작성하게 되며 LI2S 내에서 처리되는 과정은 (그림 8)과 같다.

```

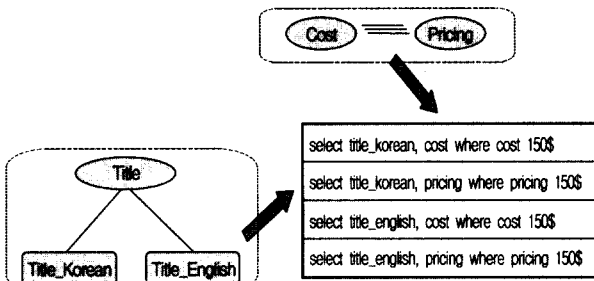
[STEP 1]: UIA(User Interface Agent)를 통해 자신이 알고있는 도메인의 개념과 용어를 가지고 질의를 작성한다. UIA는 작성된 질의를 QMA(Query Managing Agent)에 전달한다.
[STEP 2]: QMA는 UIA의 요청을 하나의 작업으로(Task)로 정의하고 시작과 종료를 독립적으로 관리한다.
[STEP 3]: QDA(Query Decomposition Agent)는 질의를 파싱하고 온톨로지에 구성되어 있는 개념들을 검사한다. 검색된 개념과 사상되어 있는 MDR의 정보를 이용해 질의를 재구성한다. 작성되어진 질의는 VM(Value Mapper)에 전달된다.
[STEP 4]: VMA(Value Mapping Agent)는 분할된 질의의 조건절에 대해 메타데이터 저장소를 참조하여 질의간 존재할 수 있는 값 이질성을 판단한다. 값 이질성이 존재할 경우, 지식베이스를 참조하여 값 이질성이 해결된 질의를 구성한다. 구성된 질의를 MDM에 전달한다.
[STEP 5]: MDMA(Metadata Mapping Agent)는 사상되어 있는 상위 데이터베이스 스키마 정보와 MDR을 참조하여 개별 데이터소스에 적용할 질의를 생성한다. 생성된 질의를 RA(Resource Agent)로 전달한다.
[STEP 6]: RA(Resource Agent)는 데이터소스에 질의를 요청하고 결과를 UIA에 전달한다.
    
```

(그림 8) 정보 통합 과정

예를 들어, SemQL을 이용하여 사용자가 자신이 알고 있는 서지정보를 검색하기 위하여 다음과 같이 가격(cost)이 '150' 달러인 서지정보의 제목과 가격에 대한 의미 질의를 요청한다고 가정하자. 다음과 같은 과정을 통해 정보는 통합되게 된다.

[STEP 1], [STEP 2] : SELECT title, cost WHERE (cost > 150 \$)

[STEP 3] : 파싱된 질의에서 추출할 수 있는 개념은 title과 cost이다.



(그림 9) 질의에서 추출된 개념에 대한 온톨로지 검색을 통한 질의분할

(그림 9)는 이렇게 추출된 개념에 대해 온톨로지를 이용한 추론을 통해 4개의 질의로 분할된 경우를 보이고 있다.

- ① SELECT title_korean, cost WHERE cost > 150\$
- ② SELECT title_korean, pricing WHERE pricing > 150\$
- ③ SELECT title_english, cost WHERE cost > 150\$
- ④ SELECT title_english, pricing WHERE pricing 150\$

[STEP 4] : VMA는 분할된 질의에 대해 메타데이터 저장소에 저장되어 있는 데이터요소를 검색하여 cost와 pricing의 사상 상태를 확인하고 <표 9>에서 제시한바 있는 지식 베이스의 표현을 통해 cost에 \$라는 전환유형을 추론하고 단위간 이질성을 해결하는 동적전환 규칙을 이용하여 달러에 대한 원화 변환 값을 적용한 질의를 다음과 같이 재구성한다.

- ① SELECT title_korean, cost WHERE cost > 150×1300\$
- ② SELECT title_korean, pricing WHERE pricing > 150\$
- ③ SELECT title_english, cost WHERE cost > 150×1300\$
- ④ SELECT title_english, pricing WHERE pricing 150\$

[STEP 5] : MDMA는 하위 데이터베이스의 스키마 정보와 각 MDR을 참조하여 관련 데이터베이스 스키

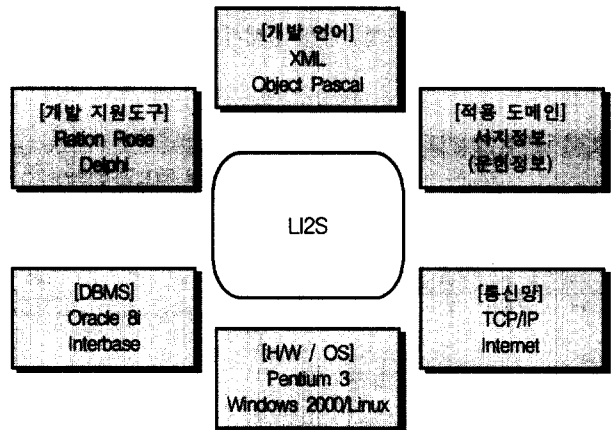
마에 맞는 질의를 다음과 같이 재구성한다.

- ① SELECT doc_kr, doc_ct FROM doc_table WHERE doc_ct > 150×1300
SELECT bibilo_title, bibilo_cost FROM bibilo_tbl WHERE bibilo_cost > 150×1300
- ② SELECT cp_cd_nm, cp_cd_ct FROM cp_cd_table WHERE comp_ct > 150
- ③ SELECT d_nm, d_ct FROM d_tbl WHERE d_ct > 150×1300
SELECT b_name, b_cost FROM b_tbl WHERE b_costt > 150×1300
- ④ SELECT bd_nm, bd_ct FROM bd_table WHERE bd_ct > 150

[STEP 6] : 이렇게 하위 데이터베이스 스키마에 맞게 구성된 질의 결과는 RA에 의해 UIA로 보내져 사용자가 통합된 정보를 볼 수 있도록 구성되게 된다.

5.4 구 현

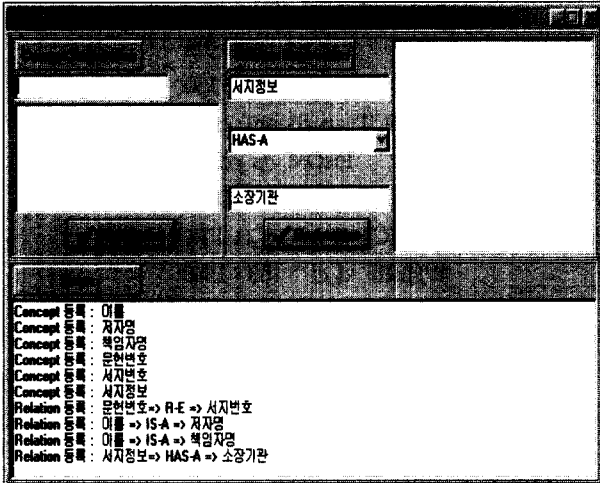
(그림 10)에서 보는 것처럼, 적용 도메인은 서지정보(문헌 정보)에 대해 메타데이터를 정의하고 개념을 추출하여 온톨로지를 구성하였다. UML 기반의 설계를 수행하였고 객체 지향 파스칼 언어와 델파이를 이용하여 오라클(ORACLE8i)과 인터베이스(InterBase)의 분산환경의 데이터베이스를 구성하였다.



(그림 10) 시스템 개발 및 구현환경

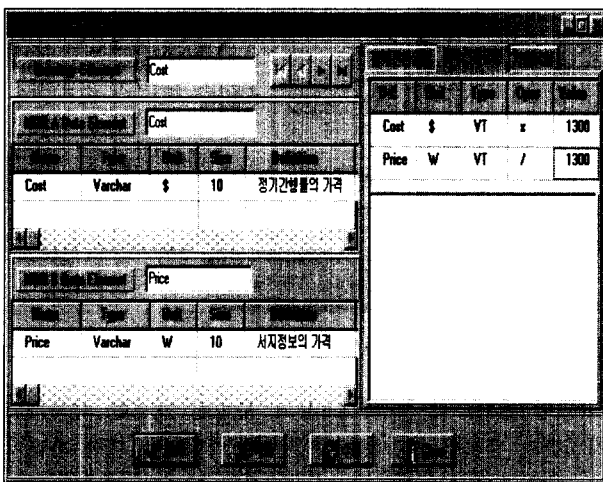
(그림 11)은 메타데이터로부터 추출된 개념을 등록하고 개념간의 관계를 설정하는 온톨로지 에디터에 대한 화면이다. 왼쪽 'Concept Registration'에서는 개념과 개념에 대한 정의를 등록하고 Relation Registration에서는 등록된 개념

들간의 관계를 등록한다. Status란에 등록된 개념과 관계에 대한 기록이 전시되어 있다. 예를 들어 '이름'과 '저자명'으로 등록된 개념이 상속관계로 사상되었음을 볼 수 있다.



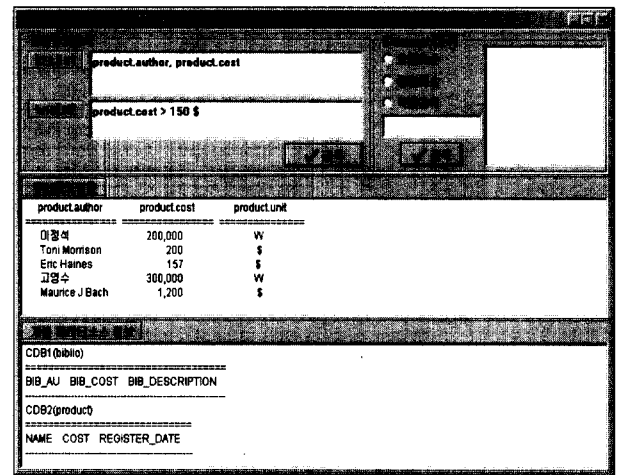
(그림 11) 온톨로지 에디터

(그림 12)는 온톨로지와 MDR 데이터 요소간 사상 관계를 검증하고 데이터 요소간 값간 이질성을 해결하기 위한 지식 베이스를 구축하기 위한 에디터 화면이다. 'Ontology Concept'에서 온톨로지에 구축되어 있는 개념을 확인할 수 있다. 'MDR A Data Element'와 'MDR B Data Element'에서 선택된 개념과 사상되어 있는 데이터요소를 확인한다. 예를 들어, 'Ontology Concept' 필드에서 'Cost'라는 개념을 검색했을 때, MDR A에는 'Cost', MDR B에는 'Price'라는 데이터 요소가 사상되어 있다. 각각의 데이터 요소가 표현하는 속성 중 'Type'은 일치하나 'Unit'이 원화와 달러화로 표현되어 데이터 이질성을 발견할 수 있다. 해당 이질성을 해결하기 위해 '동적전환유형' 탭을 선택하여 변환 연산자와 값을 설정하고 등록한다.



(그림 12) 지식베이스 에디터

이렇게 구축된 온톨로지, 메타데이터 저장소와 지식 베이스는 사용자의 질의를 재구성하는데 적용된다. (그림 13)에 서처럼 '통합정보검색'란에서 자신이 알고 있는 개념을 이용하여 질의를 수행한다. 'Concept검색'에서는 구성되어 있는 개념에 대해 세 가지 방식으로 검색할 수 있다. '통합질의결과'는 5.3절에서 기술한 정보 통합 과정을 통해 산출된 질의 결과를 전시한다. '개별 데이터소스정보'에서는 통합과정에 사용된 하위의 데이터베이스 테이블에 대한 정보를 보여준다. (그림 13)의 예는 사용자가 자신이 알고 있는 개념으로 저지정보에 대해 가격이 150달러 이상인 저자의 이름과 가격을 질의할 때의 질의문과 각 결과 값을 나타낸다.



(그림 13) 통합정보 관리자

6. 평 가

멀티 데이터베이스 시스템에서 의미 이질성을 해결하고 정보를 통합하려는 방법은 다양하게 시도되어 왔다. 전역스키마를 구성하여 정보를 통합하려는 초기의 시도는 지역 데이터베이스가 새로 추가되거나 수정되면 전역스키마의 갱신이 요구되는 정적인 단점이 있다. 본 논문에서 제안한 방법은 전역스키마를 구성하지 않기 때문에 전역 스키마의 구성에 따른 부담이 없다. 멀티 데이터베이스 언어를 지원하는 시스템에서는 전역스키마를 유지하지 않으며 사전에 데이터베이스 스키마 정보에 대해 알 필요가 없지만 분산된 데이터베이스에 대한 위치 투명성을 제공하지 않으며 사용자가 일관성을 유지해야 하는 단점이 있다. 본 논문에서는 질의에 응한 데이터베이스의 식별정보를 제공하여 사용자가 개별 데이터베이스에 직접 접근 가능하도록 MDM을 구현하였다. 다음으로 제안된 방법들은 에이전트, 도메인 온톨로지, 지능형 중재기, 그리고 고수준 질의어와 같은 다양한 기술을 이용하여 동적인 환경에 적용하도록 제안되었다. 제안된 시스템은 Ariadne, HERMES, InfoSleuth, Information Manifold, 그리고 Whirl 등이 있다. 이러한 시스템들이 공통적

으로 가지고 있는 가정은 사용자가 정보를 통합하기 위하여 필요한 도메인의 지식을 알고 있어야 한다는 것이다. 본 논문에서는 범용의 온톨로지를 구성하여 공통의 도메인 지식을 제공함으로써 사용자의 부담을 감소 시켰다.

특히, MDR 기반 정보 통합 시스템은 표준 데이터요소에 대한 생성 절차, 등록 및 인증절차를 가지고 있으며 표준으로 등록된 데이터요소를 참조하여 일관된 데이터관리가 가능하다. 그러나 범용성 측면에서는 특정 도메인에 제한적인 데이터를 표현하고 데이터 요소간 연관성을 표현하는 것에 대해 제약을 초래한다[16].

온톨로지 기반 정보 통합 시스템은 도메인에 제한적이지 않은 범용적인 의미 정보를 구성하는 연관성을 표현하여 의미망 구축이 가능하고 용어간 다양한 관계를 설정할 수 있다. 그러나 데이터 레벨의 표현 제공 측면에서 MDR이 가지는 표준적인 표현 방법을 수용하지 못한다. 확장성 측면에서는 신규 데이터베이스 추가 및 삭제에 대해 구성되어 있는 온톨로지를 변경시키는데 대한 오버헤드를 가져올 수 있다[21].

효과적인 정보 통합을 위해, 온톨로지를 이용한 정보 통합 시스템과 MDR을 이용한 정보 통합 시스템의 단점을 보완하고 장점을 수용한다. 즉, 상위 의미통합은 온톨로지를 사용하여 MDR의 데이터요소들 간의 관계를 연관지어 구성하고 하위 데이터베이스의 표준적인 메타정보는 MDR의 데이터요소 표현을 통해서 구현한다.

이렇게 두 시스템의 특징을 결합한 정보 통합 시스템(LI2S)은 MDR의 표준성을 활용한 시스템 확장적인 측면을 수용하고 온톨로지가 가지고 있는 도메인에 제한 받지 않는 범용적인 접근과 용어간 연관성을 구성할 수 있다. 또한 새로운 데이터베이스의 추가 및 삭제에 대한 오버헤드와 운영환경에 종속되지 않고 통합된 정보를 검색하고자 하는 사용자에게 범용적인 의미 질의를 수행하는 방법을 제공하게 된다.

7. 결 론

본 논문에서는 다양한 멀티 데이터베이스 환경에서 여러 데이터베이스의 정보를 통합하기 위해 기존에 제시된 정보 통합 방법론들의 장·단점을 파악하고 MDR을 활용한 정보 통합 시스템과 온톨로지를 활용한 정보 통합 시스템의 특징을 결합한 3계층 정보 통합 시스템을 제안하였다. 이렇게 제안된 아키텍처의 정보 통합 과정을 계층별 단계적으로 설명하고 MDR 사례 데이터베이스를 구축하여 LI2S를 구현하였다. 온톨로지는 사용자가 도메인의 지식 없이 자신이 알고 있는 개념을 가지고 의미 질의어를 사용하여 질의를 할 때, 의미망으로서의 질의 개념 분석과 자동적인 MDR 데이터요소 사상을 위한 역할을 담당한다. MDR은 표준이 되는 데이터요소의 등록, 관리를 통해 하위 데이터베이스의 통합 및 이질성 해결을 위한 역할을 담당한다. 계층 간 독립성은

LI2S 시스템에 새로운 데이터베이스의 동적인 확장을 위해 요구되었고 자율적이고 지능적인 에이전트를 도입하였다. LI2S의 특징은 하위 데이터베이스 스키마 정보를 알지 못하더라도 사용자가 원하는 정보를 통합할 수 있다는 데 있다. 또한 시스템에서 새로운 데이터소스를 추가하여 확장하는데 있어서도 MDR이 가진 표준성을 이용하고 에이전트의 도입을 통하여 유연성을 제공할 수 있다.

향후 연구과제로는 온톨로지 구성에 대한 자동화 방안과 VM(Value Mapper)에서 값간 사상 유형 및 규칙을 세분화하여 좀 더 신뢰성 있는 지식베이스를 구성하기 위한 방법에 대한 연구가 필요하다.

참 고 문 헌

- [1] Maurizio Panti, Luca Spalazzi, Alberto Giretti, "A Case-Based Approach to Information Integration," Proceedings of the 26th VLDB conference, 2000.
- [2] ISO/IEC, "ISO/IEC FDIS 11179 : Information technology-Specification and standardization of data elements," 1999.
- [3] Mike Uschold and Michael Gruninger, "Ontologies : Principles, Methods and Applications," Knowledge Engineering Review, 1996.
- [4] 이정욱, "멀티 데이터베이스 시스템의 정보공유를 위한 개념 기반 의미망", 고려대학교 대학원 컴퓨터학과 박사학위논문, 2001.
- [5] V. Dayal, H. Hwang, "View Definition and Generalization for Database Integration in a Multidatabase System," IEEE Transaction on Software Engineering, 1984.
- [6] R. Ahmed, P. De Smedt, W. Du, W. Kent, M. A. Ketabchi, W. A. Litwin, A. Rafli and M. C. Shan, "The Pegasus heterogenous multidatabase system," IEEE Computer, 1991.
- [7] J. Hammer, H. H. Garcia-Molina, K. Ireland, Y. Papakonstantinou, J. Ullman, J. Widom, "Information translation, mediation, and mosaic-based browsing in the tsmis system," In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1995.
- [8] C. Collet, M. Huhns, W. Shen "Resource Integration Using a Large Knowledge Base in Carnot," IEEE Computer, 1991.
- [9] A. R. Hurson, M. W. Bright, "Global Information Access for Microcomputers," Journal of Mini and Micro Computer Applications, 1991.
- [10] C. A. Knoblock et al., "Modeling Web Sources for Information Integration," In Proceedings of 11th Nat'l Conference on Artificial Intelligence, 1998.
- [11] S. Adali, K. S. Candan, Y. Papakonstantinou and V. S. Subrahmanian, "Query caching and optimization in distributed mediator systems," In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1996.

- [12] W. W. Cohen, "Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity," In Proceedings of ACM Sigmod-98, 1998.
- [13] Marian Nodine, Jerry Fowler, Brad Perry, "An Overview of Active Information Gathering in InfoSleuth," InfoSleuth Group, 1998.
- [14] Won Kim and Jungyun Seo, "Classifying Schematic and Data Heterogeneity in Multidatabase Systems," IEEE Computer, 1991.
- [15] Thomas R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," International Journal of Human-Computer Studies, 1995.
- [16] 송치양, 임성빈, 백두권, 김철홍, "컴포넌트 유통환경을 위한 컴포넌트 메타데이터 레지스트리 구축 : C_MDR", 정보과학회논문지, 2001.
- [17] Joseph P. Bigus, Jennifer Bigus, 'Constructing Intelligent agents with Java,' Wiley Computer Publishing, 1998.
- [18] 김승훈, "데이터 레지스트리의 데이터요소 값간 사상", 고려대학교 대학원 컴퓨터학과 석사학위논문, 1999.
- [19] Jeong-Oog Lee, Doo-Kwon Baik, "SemQL : A Semantic Query Language for Multidatabase Systems," In Proceedings of the 8th International Conference on Information and Knowledge Management, 1999.
- [20] 백두권, 정동원, 유상훈, 신동길, "과학기술 지식정보 공유 및 표준화 관리 모델 개발연구", 한국과학기술정보연구원, 2002.
- [21] Asuncion Gomez-Perez, "Knowledge Sharing and Reuse : Methodologies," Tutorial on Ontological Engineering : IJCAI '99, 1999.



백 두 권

e-mail : baik@software.korea.ac.kr
 1974년 고려대학교 수학과(학사)
 1977년 고려대학교 대학원 산업공학과(석사)
 1983년 Wayne State Univ. 전산학(석사)
 1985년 Wayne State Univ. 전산학(박사)

1986년~현재 고려대학교 컴퓨터학과 교수
 1989년~현재 한국정보과학회 이사/평의원
 1991년~현재 한국시물레이션학회 이사/회장
 1991년~현재 ISO/IEC JTC1/SC32 국내위원회 위원장
 2002년~현재 고려대학교 정보통신대학 학장
 관심분야 : 데이터베이스, 소프트웨어 공학, 시물레이션, 메타데이터, 정보 통합, 정보 보호 등



최 요 한

e-mail : dino9612@software.korea.ac.kr
 1994년 공군사관학교 전산학과(학사)
 2001년~현재 고려대학교 대학원 컴퓨터학과 석사과정
 관심분야 : 정보 통합, 에이전트 모델링, 분산컴퓨팅 등



박 성 공

e-mail : skpark@software.korea.ac.kr
 1996년 고려대학교 컴퓨터학과 학사
 1998년 고려대학교 컴퓨터학과 석사
 1998년~현재 고려대학교 컴퓨터학과 박사과정 재학중
 관심분야 : 데이터 엔지니어링, 이질적 데이터 통합방법, 메타데이터 레지스트리



이 정 욱

e-mail : ljo@kku.ac.kr
 1992년 고려대학교 컴퓨터학과(학사)
 1994년 고려대학교 대학원 컴퓨터학과(석사)
 2001년 고려대학교 대학원 컴퓨터학과(박사)

2002년~현재 건국대학교 자연과학대학 컴퓨터·응용과학부 조교수
 관심분야 : 정보 통합, 인공지능, 데이터베이스, 정보검색 등



정 동 원

e-mail : withimp@software.korea.ac.kr
 1997년 군산대학교 전산과(학사)
 1998년 한국전자통신연구원(위촉 연구원)
 1999년 충북대학교 전산과(석사)
 1999년 ICU 부설 한국정보통신교육원 전임강사

2000년 (주)지구넷 선임연구원
 2000년~현재 고려대학교 대학원 컴퓨터학과 박사과정
 2001년~현재 (주)라임미디어 테크놀로지 연구원
 2002년~현재 TTA TC 08 SG 02 특별위원
 관심분야 : 분산 컴퓨팅, 이동 에이전트 시스템 성능 분석, 에이전트 모델링, 메타데이터 레지스트리 기반의 정보 통합 방법론, 정보 표준화 등