데이터 표준화와 메타데이터 레지스트리 (MDR:MetaData Registry)

백두권
TTA 데이터기술위원회(TC08) 의장
고려대학교 컴퓨터학과 정교수

1. 서론

정보통신기술의 급속한 발달과 인터넷 환경의 급속한 확산에 따라 정보통신 분야에는 새로운 요구가 발생하고 있다. 새로운 비즈니스 환경에 적응하기 위해 기존의 정보통신 시스템의 통합과, 아울러 다양한 분야의 정보통신 시스템 사이의 정보공유에 대한 요구이다.

이러한 요구에 부응하기 위해 학계, 업계에서는 다양한 방법론을 개발하였고, 이를 적용하여 시스템의 통합과 정보공유를 해결하고자 노력하고 있다. 그러나, 정보통신 시스템의 통합과, 정보공유의 가장 큰 걸림돌은 시스템적인 통합 메커니즘의 개발 뿐만 아니라, 데이터의 의미(semantic), 구문(syntax), 표현(representation)의 불일치를 해결해야 한다.

정보공유를 위해서는 데이터의 표준화가 필연적이다. 정보공유를 위한 데이터 표준화를 지원하기 위한 작업의 일환으로 ISO/IEC JTC1 WG2 - 데이터 관리 및 교환(Data Management and Interchange)에서는 데이터 공유 및 교환을 위한 근본적인 해결방안으로 메타데이터 레지스트리(Metadata Registry : MDR)에 대한 표준화를 진행하고 있다. ISO/IEC 11179 - "메타데이터 레지스트리"에서 데이터의 의미, 구문, 표현을 표준화할 수 있는 프레임워크를 제시하고 있다.

메타데이터 레지스트리는 메타데이터의 등록과 인증을 통하여 표준화된 메타데이터를 유지·관리하며, 메타데이터의 명세와 의미의 공유를 목적으로 한다. 현재 여러 프로젝트에서 메타데이터 레지스트리 시스템을 구현하고 있으며, 메타데이터의 등록을 통해 각 프로젝트에서 원하는 데이터의 호환성을 유지하는 방안을 제시한다.

2. 메타데이터 레지스트리(MDR)

효율적인 정보의 교환은 보내는 사람이 의도하는 것과 동일하게 받는 사람이 정보를 번역하는 것을 보장하는 환경에서 이루어지며, 쉽게 정보의 위치를 지정해야 하고 검색되어야 한다. 따라서 정보의 의미와 표현방법 등 정보의 공유를 위해서는 통신하는 상대방과의 일정한 약속이 있어야 가능하다. 이러한 정보의 공유가 가능하도록 표준화된 의미와 형태를 가진 정보의 요소를 표준 데이터 요소라 하며, 표준 데이터 요소는 자동화된 정보처리 시스템에서 사용될 수 있다.

ISO/IEC 11179에서는 데이터를 이해할 수 있고 공유할 수 있도록 만들기 위한 표준화와 등록에 대한 내용을 설명하고 있다. 이 국제표준에서는 데이터 요소의 표준화와 등록을 이용하여 기존의 데이터 관리 방법론들에 비하여 훨씬 적은 시간과 노력으로 공유 데이터 환경을 생성할 수 있는 방법을 제시한다.

ISO/IEC 11179는 모두 6개의 부로 이루어져 있으며 각각의 내용은 다음과 같다.

(1) 제1부 "데이터 요소의 명세와 표준화를 위한 프레임워크" : 데이터 요소의 개념
(2) 제2부 "데이터 요소의 분류" : 데이터 요소 개념과 데이터 요소를 분류방법에 의하여 분류하는 절차
(3) 제3부 "MDR 메타모델" : 메타데이터 레지스트리의 메타모델
(4) 제4부 "데이터 정의의 공식화를 위한 규칙과 안내" : 명확하게 데이터 요소를 정의하는 지침
(5) 제5부 "데이터요소를 위한 명명과 식별의 원칙" : 데이터 요소의 식별에 대한 지침
(6) 제6부 "데이터 요소의 등록" : 중앙등록위원회를 통하여 데이터 요소를 등록하는 방법과 유일한 데이터 요소에 대한 식별자 부여 방법

가. 메타데이터 레지스트리의 필요성

데이터 요소들에 대한 메타데이터는 메타데이터 레지스트리에 저장된다. 메타데이터 레지스트리는 데이터의 명확한 설명을 통하여 데이터의 표준화를 지원하고, 데이터의 공유를 가능하게 한다. 등록(registration)은 데이터 공유를 지원하기 위하여 메타데이터를 기록하는 작업이다. 이러한 기록작업은 의미(semantic value)를 명확히 하기 위하여 데이터 요소단계에서 수행되어져야만 한다. 메타데이터 레지스트리는 최종 사용자(end user)가 데이터 요소의 의미를 정확하고 모호하지 않게 이해할 수 있도록 한다.

메타데이터 레지스트리는 데이터에 대한 사용자와 관리자들을 위하여 데이터를 공유하는데 필요한 기본적인 데이터 요소 특징들에 따라 메타데이터를 저장한다. 또한 식별자(identifiers), 정의(definition), 분류체계(classification categories)에 따라 메타데이터를 설명한다. 메타데이터 레지스트리는 사용자에게 데이터에 대한 정확한 의미를 발견할 수 있도록 한다. 또한, 사용자들이 데이터베이스로부터 데이터 값들을 검색하고자 한다면 원하는 데이터의 타입을 식별할 수 있도록 한다.

메타데이터 레지스트리의 기능은 다음과 같다.

- 데이터의 요청과 등록을 유용하게 한다.
- 데이터의 접근과 사용을 촉진한다.
- 메타데이터에 의하여 기술된 특징을 이용하여 데이터의 조작을 가능하게 함으로써 지능적 소프트웨어에 의한 데이터 조작을 용이하게 한다.
- CASE툴 들과 저장소(repository)를 위한 데이터 표현 메타모델의 개발을 가능하게 한다.
- 전자 데이터 교환과 데이터 공유를 유용하게 한다.

나. 메타데이터 레지스트리의 데이터 요소

메타데이터 레지스트리에 의해 관리되는 데이터의 기본단위는 데이터 요소이다.

데이터 요소는 3가지 구성요소로 이루어진다.

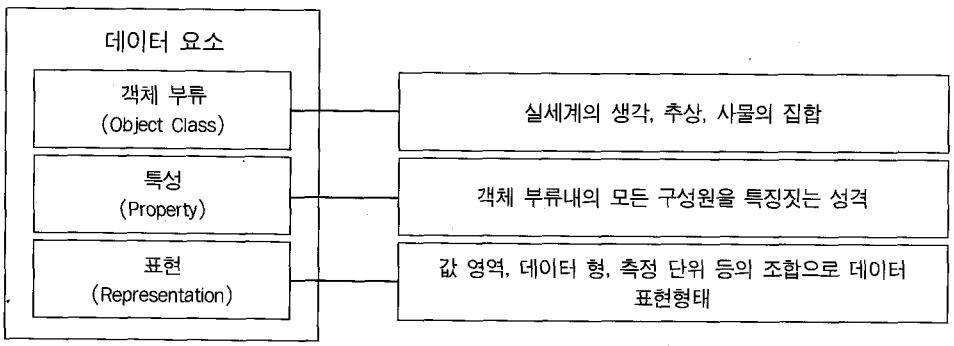
- Object Class(객체 분류) : 실세계의 생각, 추상 또는 사물들의 집합으로 명백한 범위와 의미 그리고 속성, 행위들이 같은 법칙에 의하여 정의된다.
- Property(특성) : 하나의 객체 클래스내의 구성원들이 가지는 일반적인 특성
- Representation(표현) : 데이터를 어떻게 표현하는가에 관한 문제로 value domain(값 영역), data type(데이터 타입) 등에 해당한다.

[그림 1]은 데이터 요소의 구성요소이다. 객체 클래스와 속성의 집합을 데이터 요소 개념이라 부른다. 이는 어떠한 특정 표현방법에 무관하게 데이터 요소들 자체를 나타내는 데이터 요소의 개념적인 모델이다.

타 모델의 형태로 명시한다.

메타데이터 레지스트리의 구조는 관리책임 부분(stewardship), 명명과 식별(naming and identification), 분류(classification), 데이터 요소 개념관리(data element concept administration), 개념영역과 값 영역의 관리(conceptual and value domain administration), 데이터 요소의 관리(data element administration)의 6개의 부분으로 구성된다.

- 관리책임 부분은 데이터 요소이거나 정의와 재사용이나 공유를 위한 명세를 요구하는 구성요소이다.(분류 스키마, 값 영역, 데이터 요소 개념, 개념적 영역, 객체 클래스, 특성) 조직사이에서 사용을 위한 관리를 요구하는 중요한 구성요소이다. 관리 구성요소는 데이터 요소, 분류 스키마, 값 영역, 규칙이나 분류된 구성요소를 위한 일반화이다.
- 명명과 식별은 관리되는 구성요소의 이름과



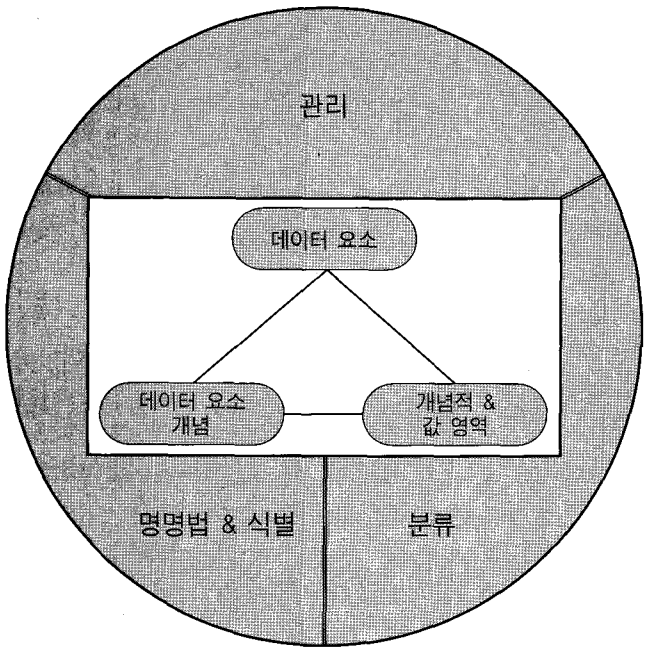
[그림 1] 데이터 요소의 구성요소

다. 메타데이터 레지스트리의 메타 모델 (Meta model)

ISO/IEC 11179 3부에서는 공유 데이터의 관리를 위한 메타 모델을 제시한다. 메타 모델은 의미적인 내용과 분산된 환경하의 사용자들이나 정보처리 시스템간의 공유되는 데이터 요소의 구문을 위한 표준과 안내를 제공하였다. 이 표준은 데이터 레지스트리의 구조를 개념적 메

그 이름의 세력범위를 제공하는 이름 문맥을 관리하기 위해 사용되는데, 고유 식별자의 부여와 데이터 요소의 이름과 관련되어 있다.

- 분류는 분류 스키마와 분류 스키마 내에 존재하는 저장소 구성요소를 관리하는 데에 사용되는데, 등록가능 요소들의 분류체계 관리 등과 관련된다.
- 데이터 요소 개념관리는 데이터 요소의 의



[그림 2] 메타데이터 레지스트리 메타 모델

미에 해당하는 개념영역에 대한 관리이다.
- 데이터 요소에 대한 관리는 데이터 요소 자체에 대한 관리이고,
- 값 영역과 개념영역의 관리는 실질적인 표현에 해당하는 부분과 관련된다.

라. 메타데이터의 등록 과정

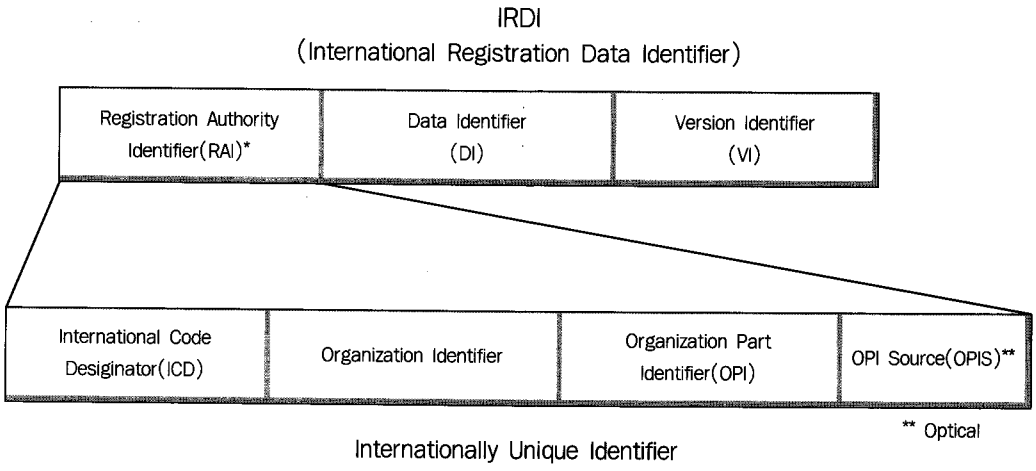
메타데이터 레지스트리에서는 중앙등록위원회를 통하여 데이터 요소를 등록하고 인증하여 데이터 표준화를 지원하게 된다.

등록된 데이터 요소의 유일성은 등록위원회 식별자, 등록위원회에서 부여한 데이터 요소의 식별자, 그리고 버전의 조합에 의하여 결정되어진다. 각각의 레지스트리는 등록 위원회에 의하여 관리된다. 다음장 [그림 3]은 등록된 데이터 요소의 부여되는 식별자에 대한 설명이다. 메타데이터 레지스트리는 만약 필요한 데이터 요소가 그 안에 존재한다면, 쉽게 접근할 수 있도록 색인하여야 한다.

등록작업은 데이터 요소를 등록하는가 마는가에 대한 내용보다는 훨씬 더 복잡하다. 물론, 좋은 데이터만을 저장하고자 하지만 실제로는 불가능하다. 그래서 등록하는 데이터의 품질을 높이기 "등록된 데이터 요소(recorded data element)", "인증된 데이터 요소(certified data element)", "표준화된 데이터 요소(standardized data element)"로 나눈다. 각 단계 사이에는 관리를 위한 상태수준이 존재한다. 이러한 상태레벨을 관리단계라고 말한다.

메타데이터 레지스트리에 저장되는 메타데이터 생명주기는 아래와 같다.

- Incomplete
  데이터 요소가 필수속성을 가지고 있지 않는 단계
- Recorded
  데이터 요소의 내용이 ISO 11179의 각 부에서 제시하는 품질 요구사항에 부합되지 않는 단계
- Certified



[그림 3] 메타데이터 식별자의 구성

데이터 요소의 내용이 ISO 11179의 각 부에서 제시하는 품질 요구사항에 부합된 단계

- Standardized
  등록기관에 의하여 인증된 데이터 요소가 만들어진 단계
- Retired
  데이터 요소가 더 이상 사용되지 않는 단계

메타데이터 레지스트리의 운영에 관련된 등록기관의 구성은 아래와 같다.

- Submitting Organization(SO)
  RA에 의하여 요구되는 기본속성과 추가적인 정보를 제공한다.
- Responsible Organization(RO)
  RA의 요청에 따라 제출된 데이터 요소의 문맥, 이름, 허용가능 값등에 대한 자문을 제공한다.
- Registration Authority(RA)
  제출된 데이터 요소에 IRDI를 부여하고, 등록된 데이터 요소를 적절히 관리한다.

3. 메타데이터 레지스트리 구현 사례

각국의 다양한 분야의 기관에서 국제표준 ISO/IEC 11179 "메타데이터 레지스트리"와 ANSI X3.285 "공유 데이터의 관리를 위한 메타모델"에 따른 레지스트리 시스템을 구현하고 있다. 이 두 표준은 데이터 요소와 메타데이터 속성을 등록하고 관리할 수 있는 시스템의 개념적인 프레임워크를 제공한다.

메타데이터 레지스트리 관련하여 가장 대표적인 구현사례로는 미국 환경청(EPA : U.S. Environmental Protection Agency)의 EDR (Environmental Data Registry), 미국 교통부(the Department of Transportation)의 ITS (Intelligent Transportation System) 데이터 레지스트리와 미국 보건복지부의 USHIK(the United States Health Information Knowledgbase) 데이터 레지스트리, 인구조사국(Census Bureau)의 인구 경제 데이터 레지스트리 등이 있고, 그외에도 호주 보건구의 NHIK(National Health Information Knowledgbase)가 있다.

가. 미국 환경청의 EDR

미국 EPA는 운영 프로그램과 다른 정부기관에 의해 정기적으로 모든 수집되는 환경 데이터에 대한 데이터 레지스트리를 개발하고 있다.

시스템 사이의 데이터를 조율하여 환경보호와 공공보건을 위한 데이터 수집, 편집 그리고 분석의 비용을 상당히 줄일 수 있다.

EPA의 EDR은 전세계적인 환경 데이터의 공유를 위해 유럽 각국의 환경 관련기관이 검토하고 있다. EPA는 ISO/IEC 11179 메타데이터 레지스트리를 선도적인 입장에서 구현하고 있다. 최신 EDR 버전에서는 EDI 메시지 집합을 등록하는 기능, 데이터 요소 값 영역을 다운로드하는 기능, 다중 정의(GILS : Global Information Locator Service 데이터 요소 등록에 필요)의 등록과 그 밖의 기능이 향상되었다. 현재 새로이 관심을 두고 추진하는 과제는 유럽 환경용어(environmental terminology) 프로그램과 함께 환경용어에 대한 문제를 다루고 있다. EPA에서는 유럽 환경용어 시스템과의 의미적인 호환성을 유지하기 위한 EDR의 서브 모듈로 용어참조 프로토타입 시스템(TRS : Terminology Reference System)을 개발중이다. 1999년 1사분기까지, EDR에는 3,601개의 데이터 요소가 저장되어있다. 2,200 데이터 요소가 환경 데이터웨어하우스를 비롯한 EPA의 시스템에서 등록되었다. EDR에는 최종 표준으로 제정된 11개의 데이터 요소, 임시 표준 24개 데이터 요소, 현재 표준으로 제정하기 위하여 검토중인 24개 데이터 요소가 있다.

최근 EPA는 환경분야의 용어에 대한 표준화에 관심을 가지고 환경용어를 제공할 수 있는 시스템을 개발하고 있다. http://www.epa.gov/edr/의 홈페이지에 EPA의 용어참조 프로토타입시스템이 운영중이다. TRS는 데이터 요소, 웹 페이지와 여러 가지 디지털 객체의 분류를 위한 용어들을 관리하기 위한 웹 기반 도구이다. 즉, ISO/IEC 11179에서 제시하는 데이터 요소의 분류를 위한 도구이다. 현재 TRS는 범용 유럽 다중언어 환경 시소러스(GEMET : General European Multilingual Environmental Thesaurus)를 저장하고 있다.

나. 미국 교통부의 ITS용 데이터 레지스트리

미국 교통부(The Department of Transportation)에서는 미국 지능형교통시스템(ITS : Intelligent Transportation System)에서 각 시스템 사이의 상호호환성(interoperability)을 보장하기 위하여 60여 종의 표준을 개발하고 있다. 표준을 개발하는 조직은 5개로 이루어져 있는데, 이 조직에서 제어용어(control vocabulary)와 정보의 다양한 접근경로(access path)를 식별하기 위한 표준을 개발하고 있다. 여러 ITS 표준 작업중에서 중요한 프로젝트는 ITS 데이터 레지스트리이다. ITS 데이터 레지스트리는 IEEE가 설계·개발하고 있고, ISO/IEC 11179를 기반으로 하고 있다.

미 교통부에서 ISO/IEC 11179를 지능형 고속도로 프로젝트에서 사용하는 직접적인 이유는 50개 주와 지역의 교통국이 ITS 시스템 내부에서 혹은 중앙정부와 ITS 외부 시스템이 가지고 있는 데이터를 의미에 따라 교환하고 상호작용할 수 있도록 하기 위해서이다.

미 교통부의 ITS 프로젝트는 ISO/IEC 11179 표준에 기반한 데이터 레지스트리를 포함한다. ITS 데이터 사전 표준을 개발하는 동안, IEEE 표준 Coordinating Commitee 32 ITS-DD/Message Set Template(MST) 작업반, DD Inter-Standard Development Organization(SDO) Coordinating 그룹과 Council of Standards Organization(CSO)에서는 ITS 데이터 레지스트리와 데이터 사전 데이터 요소의 기능적인 분석이 필요함을 인식하게 되었다. ITS 데이터 레지스트리는 IEEE P1489 - ITS를 위한 데이터 사전 표준 - 에 정의된 ITS 데이터 레지스트리에 대한 기술적인 요구사항을 만족하도록 설계되었다. IEEE P1489와 ISO/IEC 11179와의 관계는 EDR의 TRS와 같이 교통 관련 데이터 요소의 분류를 위한 목적으로 사용되고 있다. ISO/IEC 11179에서는 데이터 요소를 분류하는 것에 대한 구체적인 방법을 제시하지 않고 있

으므로 적용 분야별로 이러한 분류체계를 도입하는 것이 일반적인 접근방법이다.

ITS 데이터 레지스트리는 모든 ITS 데이터 요소와 전국의 ITS 분야를 위해 정형 명세되고 정립된 다른 데이터 개념을 위한 중앙 집중적인 데이터 사전이나 저장소(repository)이다. ITS 데이터 레지스트리는 전국의 ITS 분야에서 공통적으로 공유될 데이터를 제공하는 것이 목적이다. ITS 데이터 레지스트리의 가장 중요한 목적은 데이터 개념의 정의를 명확히 기록하여 데이터의 명확한 교환과 ITS의 기능적인 분야(즉, ITS 서브시스템과 응용시스템 사이의) 사이의 데이터와 데이터 개념(data concept)의 재사용을 지원하는 것이다.

다. 미 국방성 건강정보지식베이스(USHIK)

미 국방성 - 건강국(HA : Health Affair)은 보건증진재단(HCFA : Health Care Financing Administration)과 공동으로 미건강정보지식베이스(USHIK: the United States Health Information Knowledgbase) 데이터 레지스트리 프로젝트를 수행하고 있다. 프로젝트의 목적은 조직들 사이의 데이터 요소의 목록을 작성하고, 조율할 수 있는 데이터 레지스트리를 사용할 수 있도록 개발하고 시범 운영하는 것이다. 데이터 레지스트리를 개발하고 구현하기 위해서 환경보호국의 시스템, 호주 건강복지부의 시스템과 DoD - HA의 건강정보자원서비스(HIRS : Health Information Resource Service)를 참조하고 있다. 데이터 레지스트리는 ISO/IEC 11179와 ANSI X3.285를 따른다. 프로젝트는 시험용으로 HIPAA(Health Insurance Portability and Accountability Act) 데이터 요소를 사용하며, 데이터 레지스트리에는 X12(American Standard) HL7, NCPDP(National Council of Prescription Drug Program), NCVHS(National Committee on Vital and Health Statistics)에서 사용하는 데이터 요소를 등록, 관리한다.

DoD HA는 군의료시스템(Military Health System) 데이터 레지스트리를 위한 소프트웨어를 사용한다. HA는 데이터 레지스트리의 내용을 각 주제별로 분류하여 관리하며 유지할 책임을 공유하기 위한 파트너쉽(partnership)을 만들어 가고 있는 중이다. HA는 COTS 벤더들이 자신의 툴에서 ISO/IEC 11179에 따르도록 하는 일에 관심을 가지고 있다.

USHIK의 목표는 데이터 레지스트리를 만들고, 내용을 채우고, 시범을 보이고, 일반적으로 사용될 수 있도록 해서 여러 조직들 사이에서의 데이터 요소들을 분류하고 조화하는 작업을 보조하고, 선택된 HIPAA 데이터 요소를 사용해서 데이터 레지스트리의 기능을 시범적으로 보여주는 것이다.

데이터에 대한 명세는 HIPAA를 지원하는 Health Industry Standard와 ANSI가 인정한 위원회들 :Accredited Standards Committee X12, National Council of Prescription Drug Program(NCPDP), Health Level Seven(HL7)의 내용이 가능한 한 변경없이 사용되도록 하였다. National Committee on Vital and Health Statistics(NCVHS)의 핵심적인 데이터들도 또한 포함되었다.

USHIK는 ISO/IEC 11179와 ANSI X3.285에 기초를 두고 있다. 이 두 표준은 EPA와 AIHW의 구현에서 참조하였으며, DoD - HIRS를 개발과 구현에서도 역시 참조하였다.

메타데이터 레지스트리의 성능에 대해서 시범을 보일 목적으로 모델에 대한 요소들의 연결과 키워드의 사용은 X12 834 - Benefit Enrollment and Maintenance(004010X095)의 처리와 구현 가이드의 데이터 요소들에 초점을 맞추고 있다. 이 요소들은 호주건강복지(AIHW)모델과 연결되어 있다. NCPDP와 HL7의 데이터 요소들도 또한 AIHW 모델과 다른 데이터 요소들처럼 연결되어 있다.

USHIK 데이터 요소 레지스트리 프로젝트는 다음의 URL에서 참조할 수 있다.

http://hmrha.hirs.osd.mil/registry

4. 결론

데이터의 공유 및 교환을 지원하기 위한 메타데이터 레지스트리의 개념, 메타데이터 레지스트리 구현 사례를 소개하였다. 메타데이터 레지스트리(MDR)는 메타데이터의 명세를 표준화하고, 중앙 등록기관을 통하여 등록, 인증하게 하여 데이터의 표준화를 포괄적으로 지원하는 시스템이다. 현재 미국의 여러 기관에서 메타데이터 레지스트리에 의한 데이터 공유 환경을 구축하고 있다. 메타데이터 레지스트리는 전자정부, 전자상거래, 사이버 교육, 데이터웨어하우스, ERP 등, 정보통신 분야에서 발생하는 데이터의 이질성 문제에 대한 해결방안으로서 지식정보사회에서는 반드시 적용되어야 하는 중요한 개념이다.

• **저자약력**

1977. 2	고려대학교 산업공학 석사
1985. 12	Wayne State Univ. 전산학 박사
1990. 12 ~ 1991. 5	미국 Arizona 대학 객원교수
1986. 3 ~ 현재	고려대학교 컴퓨터학과 교수
1991. 9. ~ 현재	ISO/IEC JTC1/SC32 국내위원회 위원장
1999. 9 ~ 현재	정보통신부 정책자문위원회 위원

참고문헌

[1] ISO/IEC IS 11179-1, Information technology - Specification and standardization of data elements -Part 1: Framework for the specification and standardization of data element

[2] ISO/IEC IS 11179-2, Information technology - Specification and standardization of data elements -Part 2: Classification for Data Element

[3] ISO/IEC IS 11179-3, Information technology - Specification and standardization of data elements -Part 3: Basic attributes of data elements

[4] ISO/IEC IS 11179-4, Information technology - Specification and standardization of data elements -Part 4: Rules and guidelines for the formulation of data definitions

[5] ISO/IEC IS 11179-5, Information technology - Specification and standardization of data elements -Part 5: Naming and identification principles for data elements

[6] ISO/IEC DIS 11179-6, Information technology - Specification and standardization of data elements -Part 6: Registration of data elements

[7] ANSI X3.285 : Metamodel for the management of shareable data