제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

MDR 기반 이질 스키마간 XML 문서 자동 변환 기법†

김진관*, 김중일**, 백두권*
*고려대학교 컴퓨터학과
**㈜ 라임미디어 테크놀러지스
e-mail : jkkim@software.korea.ac.kr

An Automated XML Document Exchange Technique between Heterogeneous Schema based on MDR

Jin-Kwan Kim*, John I. Kim**, Doo-Kwon Baik*
*Dept. of Computer Science & Engineering, Korea University
**Lime Media Technologies Co., LTD.

요 약

지금까지의 정보 시스템들이 기종에 상관없이 프로그램이 실행되고 서로 데이터를 주고받을 수 있는 환경이 성취되어 왔음에도 불구하고, 데이터 자체에 대한 불일치는 범세계적인 지식 컨텐츠 공유 뿐만 아니라 전자 상거래 등의 걸림돌이 되어 왔다. 본 논문에서는 XML 문서를 통한 데이터 공유시 효과적인 상호 운용성을 얻기 위해 메타데이터 레지스트리(MDR)를 통한 XML 문서 자동 변환 기법을 제안하고자 한다.

1. 서론

지금까지의 정보 시스템들은 각각 나름대로의 목표와 용도에 맞춰 제작 및 보급 되어 왔다. 또한 인터넷의 등장과 더불어 각기 다른 시스템간의 상호운영이 중요시 되어 Java 와 같은 플랫폼 독립적인 애플리케이션 실행 환경이 도입 되었다. 하지만, 기종에 상관없이 프로그램이 실행되고 서로 데이터를 주고받을 수 있는 환경이 성취되어 왔음에도 불구하고, 데이터 자체에 대한 불일치는 범세계적인 지식 컨텐츠 공유 뿐만 아니라 전자 상거래 등의 걸림돌이 되어 왔다.

이러한 데이터의 불일치는 지금까지 표준화의 맥락에서 다루어지던 분야이다. 서지 정보의 예를 보면, MARC, UNIMARC, MARC21, KOMARC, Dublin Core, ONIX 등과 같은 수많은 표준들이 제정되어 왔고, 이들 중 상당부분이 제대로 활용되기 전에 시대의 흐름에 뒤떨어져 사장되거나 변경되어 " 표준간의 불일치" 라는 또 다른 데이터 충돌을 야기했다.

이러한 현상은 비단 서지 정보뿐만 아니라 물리, 화학, 생물, 행정 등 모든 분야에 걸쳐 나타나고 있다. 전자 상거래 분야를 보면, EDI(Electronic Document Interchange) 전자 거래 문서의 표준화로부터 시작하여, RogettaNet, BizTalk, 최근의 ebXML 에 이르기 까지, 서로 다른 정보 시스템 간의 데이터를 효율적으로 교환하기 위한 프레임워크 들이 재정되고 있다.

메타데이터란 데이터에 대한 데이터란 의미로써, 데이터를 식별하고, 활용하고, 응용하기 위한 가장 기본적인 정보의 단위이다. 앞서 설명한 데이터의 불일치, 충돌과 같은 현상은 근본적으로 메타데이터의 잘못된 생성 및 활용에 기인한다고 볼 수 있다. 1 회성 표준 제정의 한계를 경험한 데이터 표준에 관계자들은 메타데이터를 기술하는 데에 가장 효율적이며, 명확한 방법을 제안하고 이를 표준으로 제정하였다. 이것이 바로 ISO/IEC 11179 이다. ISO/IEC 11179 는

† 본 논문은 정통부 " 지식컨텐츠 기술을 위한 메타데이터 관리시스템의 개발" 사업의 일부로 수행된 결과임

기존의 다른 데이터에 대한 표준과 달리 메타데이터를 명명하고, 식별하고, 관리하는 방법론에 대한 권고안이며, MDR(Metadata Registy)라는 메타데이터 관리 시스템을 활용하여, 메타데이터의 식별 및 명명, 관리를 행할 수 있도록 제안하고 있다.

데이터의 상호 운용성은 메타데이터를 얼마나 효과적으로 공유할 수 있느냐가 관건이다. 본 논문에서는 여러 시스템 간 또는 표준간의 데이터를 XML 메시지로 메타데이터를 전달함으로써, 데이터를 담고 있는 XML 문서의 구조를 자동으로 변환하는 기법에 대해 제안하고자 한다.

2. 관련연구

2.1 X-MAP 시스템

2000 년에 David Wang 이 제안한 시스템으로 XML 로 데이터를 교환하는 다중의 이질 시스템간의 상호 운용성 조정을 위해 스키마 요소간의 의미를 자동으로 변환시켜주는 시스템이다.

X-MAP 시스템은 데이터가 가지고 있는 의미를 동등성(equal)을 정의를 이용하여 스키마 요소간 가능한 연관을 만들어 주는 것으로써, 데이터간의 이질성을 Heuristic 한 방법으로 정의하고 이런 정의를 통해서 생성된 relation 을 기반으로 Mediator 를 통해서 Domain A 에 속한 XML 문서를 Domain B 의 XML 문서로 변환한다. [1]

X-MAP 시스템은 structural Analysis, Known-relations Analysis, language/synonym Analysi 등을 자동화 하고는 있으나 Human-Assisted Analysis 가 있어야만 완전해 질 수 있는 반자동 시스템이라 할 수 있다. 또한 Mediator 를 통한 변환 시에도 Human-Assistance 가 필요하므로 변환시 마다 사람이 계속 관여해야만 한다.

2.2 11179 기반 MDR 을 이용한 표준 메타정보 공유

실세계에 존재하는 데이터들은 같은 의미에 대한 다양한 표현으로 존재한다. 이러한 다양성은 분산 환경에서의 데이터 교환 및 통합의 가장 근본적인 문제점을 야기한다.

메타데이터의 의미 불일치는 정보 시스템의 구성 단계 부터 통합에 이르는 모든 분야에서 거론되고 있으며, 이러한 불일치의 해소를 위해 ISO/IEC 에서는 메타데이터 의미 및 명칭, 관리 등에 대한 표준을 정의하고 있다. (ISO/IEC 11179 MDR - Metadata Registry)[3]

Christophe Blanchi 과 Jason Petrone 은 Metadata Registry 를 통해서 분산된 환경에서의 데이터의 상호 운용성을 높이는 방안에 대한 연구를 하였다. 이들은 Digital Object 를 통해서 메타데이터를 공유하는 방안을 제시하였다.

2.3 데이터 불일치 분류

데이터의 교환을 위해서는 데이터에 대한 정보가 필요하다. 데이터에 대한 정확한 정보는 데이터가 가지고 있는 의미, 표현, 그리고 메타데이터가 가지고 있는 구조를 파악함으로써 얻을 수 있다. 메타데이터가 가지고 있는 구조 정보가 필요한 이유는 정보를 교환, 공유, 통합하고자 하는 각각의 로컬 데이터베이스 상의 테이블구조와 구성이 모두 상이하기 때문이다.

XML 문서로 데이터를 전달할 경우, 로컬 데이터베이스의 테이블에서 표현하고 있는 스키마의 속성들의 구조대로 XML 문서를 작성하게 된다면, 다른 로컬 데이터베이스에서는 그 구조를 이해할 수 없으므로 올바른 데이터를 추출해 낼 수가 없다. 따라서 3 가지 관점에서 데이터의 이질성을 파악하여야 하며, 데이터의 교환시에 발생하는 문제점들을 해결하기 위해서는 데이터 이질성에 대한 분류가 선행되어야 한다. [표 1]은 XML 문서로 데이터를 전달할 때 MDR 을 이용하여 등록된 표준 요소들과 로컬 데이터베이스의 데이터 정보 간에 일어날 수 있는 데이터 이질성을 데이터 표현 3 요소에 근거하여 분류하였다.[4]

Data Heterogeneity		Definition
Semantic	Substitutable*	하나의 엘리먼트가 다른 엘리먼트로 대체된다.
Schematic	Composition	하나의 엘리먼트가 다른 엘리먼트들의 집합으로 구성된다.
	Decomposition	하나의 엘리먼트가 다른 엘리먼트를 구성하는 요소들 중 하나이다.
	Rearrange	하나의 엘리먼트가 다른 엘리먼트와 구조를 이루는 순서가 다르다.
Representation	MeasurementUnit*	하나의 엘리먼트가 다른 엘리먼트와 측정 단위가 다르다.
	Enumeration*	하나의 엘리먼트가 다른 엘리먼트와 다른 코드셀을 가진다.

* 표시는 MDR 만으로 해결할 수 있는 데이터 이질성 항목을 나타냄

[표 1] 데이터 이질성 분류

3. MDR_XML

ISO/IEC 11179 에서는 메타데이터의 명명, 식별, 관리에 관한 권고안을 제공하고는 있으나 실제 응용 프로그램에 사용되기 위한 방법론은 정의하고 있지 않다. 본 논문에서는 이렇게 구축된 MDR 을 활용하는 방법론으로 MDR_XML 을 정의하고 사용한다.

MDR_XML 은 다음과 같은 활용 면에서의 필요성을 가지고 있다.

첫째, Metadata 를 이용하여 명세된 MDR 내부의 데이터 요소에 의미와 표현 방식 표준 명세의 필요성

둘째, MDR 내부의 데이터 요소에 대한 구조적 표현에 대한 정의 필요성

셋째,MDR 과 거래를 원하는 각 Local Database Schema 와의 mapping rule 을 제공하기 위한 표준 필요성

MDR_XML 은 11179 기반 MDR 의 스키마를 반영하며, MDR 에 등록되어 있는 표준 데이터 요소에 대한 검색 및 이용을 가능하도록 표준 데이터 요소에 대한 정보를 출판하는 역할을 한다.

MDR_XML 은 다음과 같은 구조를 가진다. MDR_XML 은

MDR_XML 문서의 루트 요소(Root Element)로서 하나의 MDR_XML 문서는 하나의 MDR_XML 요소를 가진다.

MDR_XML 은 MSDL 작성 시 용이하게 데이터 요소에 관한 정보를 제공하기 위하여 데이터 요소를 나타내기 위해 필요한 값 영역(ValueDomain)정보인 속성(Property), 객체클래스(ObjectClass), 데이터 타입과 함께 부가적인 정보로서 데이터 요소가 사용하는 코드셀과 측정단위등을 제공한다. MDR 은 데이터 요소의 정확한 정의와 체계적인 관리를 위해서 데이터 요소 개념(Concept)과 데이터 타입, 측정단위, 코드셀의 분류(Category)를 가지고 있으므로 이들 개념(Concept)과 분류(Category)에 대한 정보도 포함하고 있다.

MDR_XML 문서는 다음과 같은 과정을 거쳐 생성한다. 먼저 데이터 요소를 검색 한 후 각 데이터 요소의 표준 단계를 확인하여 표준인 데이터 요소 및 요소 표현 정보를 활용하여 MDR_XML 문서를 생성한다. 생성된 문서에 대한 검증이 끝나면 인터넷에 게시 또는 배포하게 된다.

4. 메타데이터 정보

4.1 메타데이터 정보의 구조

메타데이터 정보는 기본적으로 MDR 에서 정의하는 데이터 요소와 실제 사용되는 데이터 간의 불일치 유형 및 변환에 필요한 정보에 대한 정의이다.

아래의 그림은 메타데이터를 나타내는 문서의 구조를 도식화한 것이다. 메타데이터 정보 문서는 표준으로 삼고자 하는 MDR_XML 의 위치와 실제 사용하고 있는 스키마의 위치를 식별할 수 있는 네임스페이스(Namespace)를 지정하는 부분과 대응정보(MAP) 각각의 데이터 요소끼리의 대응을 표현하는 부분, 이들의 변환에 필요한 정보를 저장하는 부분으로 되어있다.

하나의 메타데이터 정보 문서는 1 개 이상의 MAP(Mapping Rule)을 가짐으로써 실제 사용되는 스키마의 표현과의 차이점을 체계적으로 기술할 수 있으며, 데이터의 공유 및 교환에 이용될 수 있다.

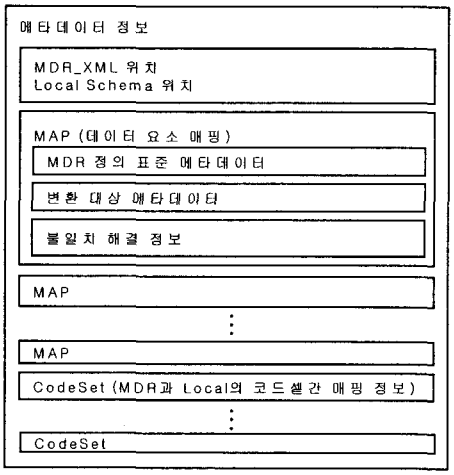
4.2 의미 정보

메타데이터 정보에는 로컬 데이터 요소와 의미상으로 일치하는 MDR 에 존재하는 표준 데이터 요소와의 관계를 표현하게 된다.

로컬 데이터 요소와 표준 데이터 요소간의 같은 의미를 가진 것을 연관시켜주는 것을 기본으로 하고, 일치하는 것이 없을 경우에는 표준요소에 새로이 등록을 요청하거나, 구조 불일치로 간주하여 여러 개의 데이터 요소를 조합하여 동등한 의미의 또 다른 데이터 요소의 집합을 만들어서 연관시키게 된다.

MDR 데이터 요소의 정보로는 표준 데이터 요소의 이름, 표준 데이터 요소의 ID, 표준 데이터 요소의 Type 정보, 그리고 수치단위를 사용하는 경우 단위정보를 기술한다.

로컬 데이터 요소의 정보로는 로컬 데이터 요소의 이름, 로컬 XML 문서에서의 xPath, 로컬 데이터 요소의 Type 정보, 그리고 수치단위를 사용하는 경우 단위 정보를 기술한다.



[그림 1] 메타데이터 정보의 구조

4.3 표현 정보

정보 통합 또는 데이터 교환 시에는 실제 시스템 상에서 사용하는 데이터 형(Type)의 불일치에 따른 정확성(Precision)의 유실이 발생한다. 가장 기본적인 데이터 형들 중 데이터의 정확도에 문제가 되는 데이터형은 정수형, 실수형, 문자형, 날짜형 등이다.[6]

이러한 표현정보는 dateType 부분에 명시한다. 표현정보에는 코드셀을 사용하는 경우 코드셀을 사용하지 않는 경우로 나뉘며 코드셀을 사용하는 경우에는 해당하는 코드셀의 집합이 어느것인지 알려주는 정보를 넣게 되며, 코드셀을 사용하지 않는 경우에는 데이터의 타입을 명시하게 된다.

4.4 구조 불일치 정보

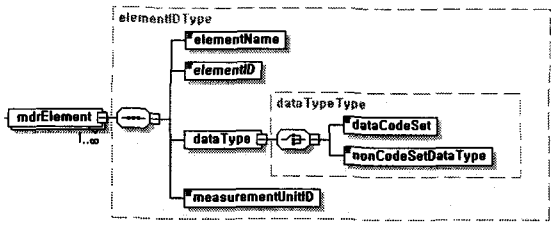
메타데이터 정보는 Substitution, Composition, Decomposition 의 3 가지 " 구조 불일치 유형" 을 정의한다.

Substitution 은 표준 데이터 요소와 로컬의 데이터 요소가 의미상 동등함을 의미하며, 해결할 불일치가 없을 경우를 말한다. 즉, 1:1 로 대체가 가능한 경우이다.

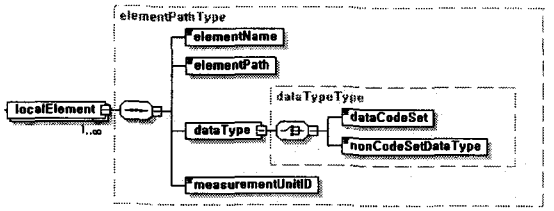
Composition 은 여러 개의 메타데이터가 하나의 메타데이터에 대응할 때 각각의 메타데이터에 의해 저장된 데이터를 분리 기호(Delimiter : eg, Comma, Semicolon, Colon, Space)를 이용하여 결합하는 경우를 말한다.

Decomposition 은 하나의 메타데이터가 여러 개의 메타데이터에 대응할 때, 하나의 메타데이터에 의해 저장된 데이터를 분리 기호를 기준으로 분리하여 각각에 대응하는 경우이다.

구조 불일치 정보에서 Rearrange 를 다루지 않는 이유는 Composition 과 Decomposition 을 표현할 때, 순서를 고려하면 따로 Rearrange 를 다룰 필요가 없기 때문이다.



[그림 2] 표준 데이터 요소의 메타데이터 정보



[그림 3] 로컬 데이터 요소의 메타데이터 정보

5. 메타데이터 정보를 이용한 문서 변환

MDR 에서는 메타데이터에 대한 의미와 표현에 대한 사항들을 정의하여 표준으로 관리할 수 있도록 한다. MDR 은 유일한 의미와 표현을 정의함으로써 데이터의 불일치를 해결하는 데에 대한 기반을 제공한다. 서로 다른 형식의 두 메타데이터가 같은 의미를 갖고 있다면, 이들의 표현상의 불일치를 해결해야 데이터의 공유 및 교환이 가능하다. 이를 위해 각각의 불일치 경우에 대한 해결 방법이 필요하며, 이를 정의하고 기술하는 것이 메타데이터 정보의 역할이다.

메타데이터 정보를 이용한 메시지 변환은 변환과 대치(Conversion & Substitute)의 반복이다. 가장 기본적인 변환 유형은 "대치가능(Substitutable)" 이다. 이는 변환할 필요 없이 그대로 교환 가능하다는 의미이다. 기본적으로 의미가 같은 모든 메타데이터들은 그 표현, 데이터 형식의 변환을 통해 대치 가능한 형태로 변형되며, 변환의 마지막 단계는 대치이다

[그림 4] 메타데이터 정보를 이용한 메시지 변환

메타데이터 정보를 이용한 로컬 A 타입의 문서에서 로컬 B 타입으로의 메시지 변환은 다음과 같은 2 가지 단계를 거침으로써 이루어지게 된다.

1 단계(표준화 단계) : 로컬 A 타입의 메시지를 표준 메시지 형태로 변환하는 단계이다.

2 단계(로컬화 단계) : 표준 메시지를 로컬 B 타입의 메시지로 변환한다.

6. 결론 및 향후 연구 방안

현재 메타데이터에 대한 연구 및 개발은 메타데이터 등록기를 기반으로 XML 관련 기술을 적용한 해결 방법이 주류를 이루고 있다. 즉 기존의 방법론이 가져온 한계를 XML 이라는 신기술과 ISO/IEC 11179 라는 메타데이터 생성, 관리 방법론을 통하여 해결하려는 시도인 것이다.

본 논문에서는 XML 문서를 통한 데이터의 교환,공유, 통합을 자유롭고 효과적으로 하기위해 메타데이터 정보를 통한 XML 문서간의 자동 변환 기법에 대해 제시하였다.

본 논문에서 제시한 방법의 활용 방안으로 현재 관련된 e-Business Framework 및 CASE TOOL Repository 로서 ISO/IEC 11179 MDR 활용방안에 대한 연구가 진행중이다.

참고문헌

[1] David Wang, Automated Semantic Correlation between Multiple Schema for Information Exchange, M.I.T., MM, May 2000

[2] Christophe Blanchi, Jason Petrone, Distributed Interoperable Metadata Registry, D-Lib Magazine, December 2001

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

[4] 김진관, 김중일, 최오훈, 백두권, MDR 을 이용한 XML DTD 이질성 해결기법, 한국정보과학회 추계학술발표 논문집(I 권), pp 67~69, 2002

[5] 최오훈, 김중일, 김진관, 백두권, MDR 기반 하이브리드 e-business 프레임워크, 한국정보과학회 추계학술발표논문집(III권), pp67~69, 2002

[6] ISO 11179 기반 데이터 레지스트리에서 데이터 요소간 값 사상, 김승훈, 박대하, 나홍석, 백두권, 한국정보과학회 춘계 학술발표논문집(B 권), pp.48~50. 1999

[7] 홍종하, 양유승, 나홍석, 백두권, " 메타데이터 레지스트리를 이용한 XML-문서 교환 방법" , 정보과학회 춘계 학술발표논문집, 2001