국가 과학기술 R&D 기반정보 온톨로지 구축 및 적용

이미경*°, 정한민*, 이승우*, 성원경*
*한국과학기술정보연구원 정보시스템연구팀
e-mail : {jerryis, jhm, swlee, wksung}@kisti.re.kr

Construction and Application of National Science and Technology R&D Reference Information Ontology

Mikyoung Lee*°, Hanmin Jung*, Seungwoo Lee*, Won-Kyung Sung*
*Information System Research Team
Korea Institute of Science and Technology Information

요 약

과학기술 연구자들의 협업을 지원하기 위해서 정보 자원 공유에 기반한 정보 유통 체제가 필요하나 현재 정보 유통 체제에서는 서로 이질적인 형태로 정보가 표현되어 있기 때문에 정보 공유의 기술적 한계를 갖고 있다. 그리고 대량의 정보 속에서 사용자가 원하는 정보를 선별하여 제공하기 위해서는 새로운 정보 유통 플랫폼이 필요하다. 본 논문에서는 지식 기반 정보 유통 플랫폼 상에서 이용되는 국가과학기술 R&D 기반정보를 지식화하기 위해 국가과학기술 R&D 기반정보 온톨로지를 구축하여 이용함으로써 각 기관별로 관리하고 있는 인력, 성과물 등의 과학기술 R&D 기반 정보의 표준화된 지식관리 체계로 이용할 수 있다. 우리는 국가과학기술 R&D 기반정보 온톨로지를 구축하기 위하여 한국과학기술정보연구원의 내부 성과물 정보를 분석하여 클래스들과 속성 관계들을 생성하였다. 또한, 한국과학기술정보연구원(KSITI) 내부 성과물 정보의 실제 데이터들을 이용하여 온톨로지의 Individuals 를 생성하였다.

정보 유통 플랫폼에서 온톨로지 형태로 구축된 지식을 이용하면 과학기술 R&D 기반정보에 대한 효율적인 관리가 가능하고, 정형화된 형태의 지식으로 개념화했기 때문에 지식 데이터의 공유와 재사용이 가능하다. 또한 단순 질의 검색이 아닌 의미 기반 추론을 이용한 지식 검색이 가능해지는 장점을 가진다. 우리가 구축한 국가 과학기술 R&D 기반정보 온톨로지를 이용하여 정보유통플랫폼(OntoFrame-K)에서 연구자 네트워크, 연구자 추적, 연구맵의 추론 서비스를 제공한다.

1. 서론

과학기술이 발달함에 따라 최단 기간 내에 최고의 연구 성과를 얻을 수 있도록 같은 분야의 연구자들간에 협력 연구를 지원할 수 있는 정보 자원 공유에 기반을 둔 협력 연구 체제를 구축할 필요성이 더욱 커지고 있다. 현재 유통 대상 정보 자원이 증가하면서 정보의 형태적 유사성에 기초한 기존의 정보 유통 체제의 기술적 한계가 드러남에 따라, 대량의 정보 자원으로부터 수요자가 원하는 정보만을 선별하여 제공할 수 있는 새로운 정보 유통 기술이 절실히 요구되고 있다. 이를 위해 본 논문에서는 차세대 정보 유통 시스템을 위해 시맨틱 웹 기술을 적용하고자 한다[1]. 선진국들은 2000 년도 초에 이미 시맨틱 웹의 기반 기술을 어느 정도 확보하고 있으며 시맨틱 웹 기술에 기반한 정보 유통 응용 시스템 개발 단계에 진입해있다. 영국의 AKT(Advanced Knowledge Technologies) 프로젝트의 CAS(CS Aktive Space)시스템은 시맨틱 웹 기

술을 응용한 대표적인 과학기술 정보유통 시스템으로 주목 받고 있다[2].

시맨틱 웹과 온톨로지를 정보유통 플랫폼에 접목시키면 의미 기반의 검색과 추론이 가능하고 서로 이질적인 지식 정보들을 온톨로지로 구축함으로써 지식의 재사용성을 높이고 지식 정보의 공유 효과를 누리게 된다. 지식 기반 정보 유통 플랫폼은 온톨로지 기반의 추론 서비스를 이용하여 연구자 간 협업에 필요한 정보와 서비스를 제공하는 역할을 한다[1][3][4].

2. 시맨틱 웹과 온톨로지

2.1. 시맨틱 웹

시맨틱 웹은 현재 우리가 사용하고 있는 웹의 확장된 개념이다. 가장 큰 차이점은 현재의 웹은 단순한 키워드나 디렉토리 방식으로 검색이 이루어지나 시맨틱 웹은 의미기반 검색 방식으로 이루어지는 점이다[5]. 시맨틱 웹은 웹 상의 정보에 잘 정의된 의미를 부여함으로써 사람뿐만 아니라 컴퓨터도 쉽게 문서의 의미를 해석할 수 있도록 하여 컴퓨터를 이용한 정보의 검색 및 해석, 통합 등의 업무를 자동화하기 위한 목적으로 제안되었다.

Tim Berners-Lee 는 시맨틱 웹이 기존의 웹과 완전히 구별되는 새로운 웹의 개념이 아니라 현재 웹을 확장하여 웹에 올라오는 정보에 잘 정의된 의미를 부여하고 이를 통해 컴퓨터와 사람이 협동적으로 작업을 수행할 수 있도록 하는 패러다임이라고 정의하였다[6]. 이러한 “잘 정의된 의미”를 다루고자 하는 것이 바로 시맨틱 웹 온톨로지 언어의 역할이다[7].

시맨틱 웹의 등장은 정보를 검색할 때 더욱 정확한 결과를 가져오게 되었고 서로 다른 이형질 정보를 통합하고 비교할 수 있게 되었으며 어떤 자원에 대해서도 의미적이고 기술적인 정보를 연관시켜 웹 서비스의 자동화를 가능하게 한다..

2.2. 온톨로지의 필요성과 역할

온톨로지에 대한 정의는 여러 가지가 있지만 Gruber 는 온톨로지는 “공유된 개념화(shared conceptualization)에 대한 정형화되고 명시적인 명세(formal and explicit specification)”라고 정의하였다[6]. 이 정의를 세부적으로 살펴보면 개념화는 사람들이 사물에 대해 생각하는 바를 추상화한 모델로 특정한 분야에 국한시켜 논의한다.

개념의 타입이나 사용상의 제약 조건들이 명시적으로 정의되어야하며, 온톨로지는 프로그램이 이해할 수 있어야 하고 여러 단계의 정형화가 존재할 수 있다. 온톨로지는 합의된 지식을 나타내므로 어느 개인에게만 국한되는 것이 아니라 그룹 구성원이 모두 동의하는 개념이다. 온톨로지는 웹 기반의 지식처리나 응용 프로그램 사이의 지식 공유, 재사용들을 가능하게 하는 아주 중요한 요소로 자리잡고 있다.

온톨로지는 개념과 관계들로 구성된 네트워크로 어느 특정 도메인에 관련된 개념들을 계층적 구조로 표현하고 추가적으로 이를 확장할 수 있는 추론규칙을 포함한다. 계층적 구조를 나타내는 계층 분류는 객체의 클래스와 서브클래스, 그들간의 관계를 정의한다. 추론 규칙은 프로그램이 새로운 사실을 자동으로 추출하거나 제약조건에 맞지 않는 오류를 찾아내는데 이용된다[8].

3. 온톨로지 구축 과정

온톨로지를 구축할 때는 구축 목적을 가장 먼저 고려해야 한다. 구축하려는 온톨로지가 소규모 단위의 시스템에서만 공유될 것인지 다양한 응용에서 재사용될 것인지 대규모 단위에서 재사용될 것인지에 따라 온톨로지의 성격이 달라지기 때문이다.

다음으로 온톨로지가 구축될 분야의 특성을 고려하여 분야의 적합성 여부를 판단하고 구축 범위를 확정한 후, 해당 분야가 가지는 공통적인 특성과 개별적인 특성을 분석하여 온톨로지 전체적인 구조를 설계한다. 이 단계는 클래스와 속성, 인스턴스를 설정하는데 필요하다.

온톨로지 내부 구조를 설계하기 위해서는 개념, 속성, 제약조건 등을 설계해야 한다. 내부 구조의 개념은 Class 로 표현하고, 클래스들의 관계와 클래스의 속성은 Property, 제약 조건은 Restriction 으로 표현한다. 이 과정에서 각각의 개념과 관계를 문맥에서 정확하게 산출하고 어떤 개념과 관련되는지 반드시 확인해야 한다.

온톨로지 내부 구조 설계 단계를 거친 후, 실제 온톨로지로 구축하기 위해서 온톨로지 구축 언어를 결정한다. 이 단계에서는 현재 구축하려고 하는 관계들의 잘 표현할 수 있는 표현력을 가진 언어를 선택하는 것이 바람직하다. 선정한 구축 언어를 이용하여 미리 설계해놓은 클래스와 속성 관계들을 구축하고 실제 인스턴스 데이터들도 구축한다. 온톨로지 구축과 병행하여 온톨로지의 유효성을 검사해야하며, 버전을 관리하여 지속적인 유지 보수가 가능하게 한다.

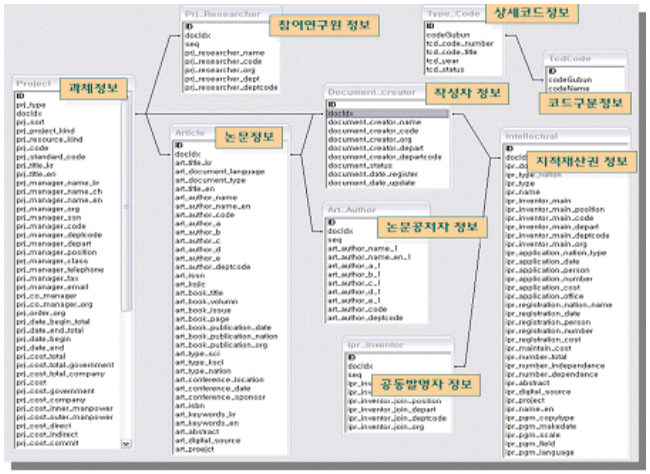
4. 국가과학기술 R&D 기반정보 온톨로지 구축

4. 1 온톨로지 구축을 위한 정보 분석

우리는 정보 유통 플랫폼에서 이용될 온톨로지를 구축하기 위해 KISTI 내부 성과물 데이터를 이용한다. 성과물 관리 데이터베이스는 과제, 참여 연구원 정보, 출판물, 공저자 정보, 지적 재산권, 공동 발명자 정보, 코드 구분, 상세 코드 구분 등으로 (그림 1)과 같은 스키마 구조를 가진다. 데이터베이스의 테이블 정보와 테이블 필드 정보를 통해 class 와 property 의 관계를 설정하는 기초 자료로 활용하였다. 과학기술 R&D 기반정보 온톨로지는 (그림 1)의 KISTI 내부 성과물 정보 스키마를 분석하여 인력정보와 과제정보를 중심으로 설계, 구축하였고, 학술정보와 과제정보도 상호 연계할 수 있도록 구축하였다.

(그림 1)의 성과물 데이터베이스에 저장된

데이터들은 온톨로지의 Individuals 로 구축되었다. 하지만 Database 데이터들은 유효하지 않은 형태를 가지거나 누락된 정보들이 많았기 때문에 모든 데이터들은 맞춤법을 교정하고 누락된 정보가 포함된 필드는 제거하는 데이터 정제 작업을 거쳤다. 그리고 DB 스키마를 이용한 정보 분석 방법은 온톨로지를 구조화하기 쉬운 장점을 가지나 DB 에 의존적인 형태의 온톨로지가 설계되는 단점을 가지고 있다.



(그림 1) KISTI 내부 성과물 정보 스키마의 예

4.2 온톨로지 설계

(1) Class 설계

기반 정보 온톨로지를 구축하기 위해서 필요한 개념들을 내부 성과 정보 스키마를 이용하여 클래스화하였다. 온톨로지의 상위 클래스는 과제(Project), 과학기술인력(Person), 출판물(Publication), 기관(Group), 지적 소유권(Ownership), 행사(Event)의 6 개의 클래스로 설계하였으며, 각 클래스들은 여러 개념의 하위 클래스들을 가지게 된다.

- Event Class: 과학 기술 분야와 관련된 행사들을 개념화한 클래스이다. 이 클래스는 Publication 클래스와 관련을 가지고 있다.
- Group Class: 과학기술 분야에 관련된 각종 집단들을 개념화한 클래스이다. Group 는 과제와 관련된 기관, 출판물과 관련된 기관, 과학기술 인력의 소속 기관, 행사 주관 기관 등 여러 클래스들과 연계성을 가지는 클래스이다.
- Ownership Class: 지적소유권 개념을 클래스화한 것이다. 디자인, 특허, 실용신안, 프로그램, 상표의 5 가지 하위 클래스를 가진다. 지적 소유권들은 과제에서 유발되는 경우가 많기 때문에 Project, Person, Group 클래스와 관련을 가진다.
- Person Class: 과학기술 분야와 관련된 각종 인력에 대한 개념을 클래스화한 것이다. Publication, Ownership, Organization 클래스와 연계성이 가장 많은 클래스이다.
- Project Class: 성과물 중에 과제와 관련된 개념을 클래스화한 것이다. 이 클래스는 Publication, Ownership, Person 클래스와 관련을 가진다.
- Publication Class: 각종 출판물에 대한 개념을 표현한 클래스이다. Event 와 Project 클래스와 관련을 가진다.

(2) Property 설계

Property 는 개념화한 클래스의 속성 정보들로 클래스들간의 관계와 속성 값, 제한 조건을 생성하기 위해 설계하였다. 여기서 사용되는 제한 조건으로는 하위 클래스가 상위 클래스의 상속 속성과 제한 조건을 비롯하여 추가적인 속성과 제한 조건 정보를 추가할 수 있도록 설계하였다.

- Event Class Properties: 행사명, 주최기관, 행사 논문 등의 일반 속성을 가지며 행사 개최일자와 장소 등에 관계 속성을 가진다.
- Group Class Properties: 그룹의 명칭, 위치 및 주소, 전화번호, 홈페이지 주소 등의 일반 속성과 수행 프로젝트, 프로젝트 투자 자금 지원, 출판물 출판, 소속 인력 등의 관계 속성을 가진다.
- Ownership Class Properties: 지적 소유권의 명칭, 발생일자, 소유자, 관련 프로젝트 등의 속성을 가진다.
- Person Class Properties: 사람의 이름, 주소, 연락처 등의 일반적인 속성과 소속 기관, 참여 프로젝트, 발표 논문, 연구 분야 등의 관계 속성을 가진다.
- Project Class Properties: 과제명, 과제 금액, 과제 기간 등의 일반 속성과 수행기관, 자금지원기관, 과제책임자, 과제유발출판물, 지적소유권, 참여연구원, 연구분야 등의 관계 속성을 가진다.
- Publication Class Properties: 논문의 제목, 저자, 키워드, 발생연도, 연구분야 등의 속성을 가진다.

4. 3 온톨로지 구축

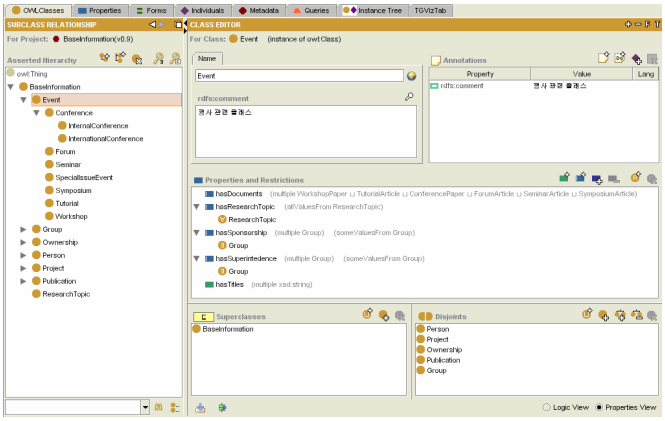
(1) 온톨로지 구축

설계된 온톨로지 개념들을 구축을 하기 위해 OWL 을 이용하였다.

Event 클래스는 과학 기술 분야와 관련된 학술대회, 포럼, 세미나, 튜토리얼, 워크숍 등의 모임이나 행사를 포함한다. Group 클래스는 과제와 관련된 기관, 출판물과 관련된 기관, 과학 기술 인력의 소속 기관, 행사 주관 기관 등 각종 기관 정보를 포함한다. Ownership 클래스는 과학 기술 분야와 관련된 각종 지적 소유권 개념을 표현한 것으로 디자인, 특허, 실용신안, 상표, 프로그램 등이 여기에 하위 클래스로 포함한다. Person 클래스는 인력의 개념을 포함한 것으로 과학 기술 분야의 전문가 추천 및 전문가 그룹(COP: Communities Of Practice) 형성에 이용된다. 이 클래스는 논문, 지적 재산권, 과제 관련 인력의 모든 정보가 포함된다. Project 클래스는 위탁과제, 협동과제, 용역과제, 주관 과제 등 과제와 관련된 정보를 포함하고 논문과 지적재산권 등 성과물과

관련이 있다. Publication 클래스는 논문관련 연구 성과물에 대한 정보를 포함하고 있다. 도서, 논문, 학술지, 학회지, 기술 보고서 등 문서 관련 정보를 포함한다.

개념화된 모든 정보를 클래스로 구축한 후 속성관계를 구축하여 Property 를 생성한다. 각 Property 는 클래스들 간의 제약 조건을 가지며 좀 더 명확한 관계를 선언한다. 본 논문에서 구축한 기반정보 온톨로지는 (그림 2)와 같다. OWL 을 이용하여 온톨로지를 구축하였으며 89 개의 Classes, 85 개의 Properties, 2,964 개의 Individuals 로 구성되어 있다.



(그림 2) 국가 과학기술 R&D 기반정보 온톨로지

(2) 인스턴스 구축

온톨로지 구조를 정의하였다면 실제 클래스들의 인스턴스인 KISTI 내부 성과물 관리 데이터베이스의 Fact Data 를 Individuals 로 구축한다. 각 클래스의 인스턴스들은 국내외 학술대회, 심포지움, 워크숍 등 160 여개, 회사, 정부기관, 연구소 등의 기관 340 여개, 특허, 프로그램을 포함한 지적재산권 260 여개, KISTI 내부 연구원과 과제에 참여한 교수, 학생을 포함한 인력들의 데이터 340 여개, 실제 내부 과제의 데이터 380 여개, 논문 자료로 1400 여 편 등의 인스턴스 데이터들을 Individauls 로 구축하였다.

5. 온톨로지 활용

본 논문에서 구축한 국가 과학기술 R&D 기반정보 온톨로지는 정보유통 플랫폼(OntoFrame-K)의 추론 서비스에 이용된다. 추론 서비스 시스템은 기반 정보 온톨로지와 추론 서비스 시나리오에서 필요한 추론 규칙을 이용하여 추론 서비스를 제공한다. 추론 서비스는 연구 사업과 학술정보 관점에서 특정 연구 토픽에 대한 연구자 네트워크를 획득하고 해당 분야의 전문가를 추천, 순위화하는 연구자 네트워크, 정해진 기간 내에서 특정 연구자의 연구사업 정보와 학술정보 등 연구자 동향을 제시하는 연구자 추적, 정해진 기간 내에서 지역이나 부서의 연구 동향 정보를 제공하는 연구맵의 3 가지 시나리오에 의한 서비스를 제공한다.

(그림 3) 연구자 추적 추론 서비스

6. 결 론

본 논문에서는 과학기술 연구자들의 협업 연구에 도움을 주기 위한 지식기반 정보유통 플랫폼에 이용되는 국가과학기술 R&D 기반정보 온톨로지를 구축하였다. 온톨로지 구축 지침에 따라 OWL 로 구축된 온톨로지는 Event, Group, Person, Project, Ownership, Publication 의 6 개의 상위 클래스를 기반으로 구성된 Classes, 클래스들간의 속성관계, 제한조건을 나타내는 Property, 인스턴스 데이터들인 Individuals 로 구성되어 있으며, 현재 구축된 온톨로지는 총 37,656 개의 트리플로 변환된다. 우리가 구축한 국가 과학기술 R&D 기반정보 온톨로지는 서비스 기반의 추론 규칙과 함께 지식기반 정보유통 플랫폼(OntoFrame-K)의 추론 서비스에서 활용된다.

참고문헌

[1]이미경, 정한민, 성원경, “지식기반 정보유통 플랫폼 개발”, 제 10 회 한국 과학기술 정보인프라 워크샵, 제 1 권 1 호, pp. 321-327. 2005.

[2] http://triplestore.aktors.org/demo/AKTiveSpace/

[3] 정한민, 성원경, 박동인, “연구자간 협업 지원 서비스를 위한 지식 베이스 설계”, 제 17 회 한글 및 한국어 정보처리학술대회, pp.173-178, 2005.

[4] 구남앙, 정한민, 김원용, 성원경, “정보 유통 플랫폼 상에서의 정보 검증 및 검색 방안”, 제 10 회 한국과학기술정보인프라워크숍, 제 1 권 2 호, pp103-109, 2005.

[5]최중민, “시맨틱 웹의 개요와 연구동향”, 정보과학회지, 제 21 권, 제 3 호, pp.4~10, 2003.

[6] T. Bernes-Lee, J.Handler, and O. Lassila, “The Semantic Web,” Scientific American, 2001.

[7] A. Gomez-Perez, O. Corcho, “Ontology Language for the Semantic Web,” IEEE Intelligent Systems, vol.17, no.1, January/February, pp.54-60, 2002.

[8]김은경, 남영준, “시맨틱웹을 위한 온톨로지 구축방법에 관한 비교 연구”, 정보관리연구, vol.35, no.2, pp.57-85, 2004.