국가과학기술 R&D 기반정보 온톨로지

Semantic Web Ontology for Research Community

강인수, 정한민, 이승우, 김평, 성원경
한국과학기술정보연구원, NTIS 사업단

Kang In-Su, Jung Han-Min, Lee Seung-Woo,
Kim Pyung, Sung Won-Kyung
KISTI

요약

시맨틱웹 분야의 온톨로지는, 분산 이종 스키마를 갖는 정보에 대한 의미적 통합, 에이전트에 의한 자동화된 유통, 그리고 명시적으로 표현되지 않은 암묵적 지식의 추론을 가능케 하는, 논리에 기반하여 잘 정의된 도메인 의존적인 콘텐츠로 볼 수 있다. 이 논문에서는 과학기술 연구분야 도메인에서 정의된 하나의 시맨틱웹 온톨로지인 '국가과학기술 R&D 기반정보' 온톨로지를 소개한다. 이 온톨로지는, 논문, 과제, 특허 등과 같은 연구성과물을 바탕으로 특정 분야 전문가들이나 연구자 네트워크를 추론해 냄으로써 실제 연구자나 연구과제 기획/평가자들에게 도움을 주기 위한 목적으로, 현재 한국과학기술정보연구원에서 개발되고 있다. 이 논문에서는 온톨로지의 스키마 구성 요소로써 클래스와 속성을 기술하고, 온톨로지에서의 인스턴스 표현의 예를 보인다. 상기 온톨로지는 인력, 기관 등과 같은 실세계 인스턴스의 고유한 신원을 온톨로지 내에 대응시키기 위해, 실세계 신원의 중의성 해소 기법과 함께 개별 클래스에 종속적인 URI할당지침을 사용하여 온톨로지 내에 인스턴스를 표현하고 있다.

Abstract

Semantic web ontologies can be viewed as logic-based domain-oriented contents which allow distributed and heterogeneous information to be semantically integrated, automatically circulated, and enable implicit knowledge to be reasoned. This paper describes the 'Science and Technology Research Area' ontology which is being developed by the Korea Institute of Science and Technology Information (KISTI). This ontology was defined to assist actual researchers and project planners to grasp the researchers community from a variety of viewpoints. We describe classes and properties as ontology components and exemplify the representation of real instances in the ontology. In order to represent the identities of real world instances within the ontology, the above ontology employs both class-dependent URI assignment schemes and the identity resolution methods.

I. 서 론

인간 중심으로 만들어진 현재의 웹 문서들은 그 내용에 대한 의미표현이 결여되어 있어 기계에 의한 자동 처리가 어려우며, 그 양이 방대해 짐에 따라 현재의 키워드 기반의 검색 기법으로는 원하는 정보를 찾는 데 한계가 드러나고 있다. 이러한 문제를 해결할 수 있는 대안으로 제안된 시멘틱웹은, 개별 응용 도메인마다 미리 정의된 온톨로지의 개념 클래스와 속성관계를 사용하여 웹문서와 같은 콘텐츠들에 의미태그를 부착함으로써 의미레벨의 콘텐츠 표현을 가능하게 한다.

시멘틱웹에서 사용하는 온톨로지는 특정 도메인에 대해 사람들 사이에 공유되는 개념체계를 명시적이고 형식적으로 규정해 놓은 것이다 [2]. 이러한 온톨로지는 분산 이종 스키마를 갖는 정보에 대한 의미적 통합, 에이전트에 의한 자동화된 유통, 그리고 명시적으로 표현되지 않은 암묵적 지식의 추론을 가능케 하는, 논리에 기반하여 잘 정의된 도메인 의존적인 콘텐츠로 볼 수 있다.

본 논문에서는 현재 한국과학기술정보연구원에서 개발 중인 '국가과학기술 R&D 기반정보' 온톨로지(이하 *기반정보 온톨로지*)를 소개한다. 기존의 온톨로지들과 달리 *기반정보 온톨로지*는 URI(Universal Resource Identifier)서버에 기반하여 인스턴스를 표현하고 있다. 온톨로지 표현에 있어 URI서버의 도입은, 클래스 종속적인 데이터타입속성들을 온톨로지 스키마에 표현하는 대신 URI서버에 둠으로써 온톨로지 스키마를 간소화시키고 인스턴스 데이터베이스를 경량화시켜 온톨로지에 기반한 추론엔진의 부담을 줄여주는 장점이 있다.

논문의 구성은 다음과 같다. 2장에서 기존연구를 다루고, 3장에서 *기반정보 온톨로지*의 개요, 설계원칙, 스키마에 대에 기술한다. 4장에서 *기반정보 온톨로지*의 URI서버 지향적 인스턴스 표현에 대해 기술한다. 5장에서 결론을 맺는다. 논문에서 온톨로지 스키마 구성요소는 **볼드체**로 표시하였다.

II. 기존 연구

기존에 개발된 연구개발 분야의 대표적인 온톨로지로는 KA2 온톨로지 [1], AKT 온톨로지[1]와 KA2 온톨로지를 확장하여 만든 SWRC 온톨로지 [3] 등이 있다. 이 외에도 KA2를 수정/확장한 Science 온톨로지[2]와, SWRC 온톨로지에 기반하여 만들어진 Bibster 온톨로지, SWRC+COIN 온톨로지 등이 있다. 이들 온톨로지들은 공통적으로 인력, 기관, 과제, 출판물 등을 온톨로지 스키마 요소로 포함하고 있다. 본 논문에서 소개하는 *기반정보 온톨로지* 역시 연구개발 분야를 다루고 있으므로 기존의 온톨로지들과 모델링 관점에서 큰 차이는 없다. 하지만, 성과물 중심으로 설계된 *기반정보 온톨로지*는 논문, 특허와 같은 성과물의 실적이 인력이나 기관에 정확히 누계될 수 있도록 하는 장치를 갖고 있다는 특징이 있다. 이러한 장치들로써 *기반정보 온톨로지*는 인력이나 기관의 인스턴스 표현에서 발생하는 다의어 및 동의어 문제를 해소하는 전처리를 거치고 있으며, 성과물의 특정 순위 저자를 모델링한 클래스를 추가로 포함하고 있다.

III. 국가과학기술 R&D 기반정보 온톨로지

1 온톨로지 개요 및 설계 원칙

*기반정보 온톨로지*는 과학기술분야의 연구/개발자들의 협업 연구를 가상 공간에서 지원하기 위한 소프트인트라의 하나로 개발되고 있다. 현재의 *기반정보 온톨로지*는 논문, 과제, 지적재산권 등의 연구/개발 성과물을 중심으로, 관련된 인력, 기관, 연구 분야/주제 등을 모델링하고 있다.

기반정보 온톨로지 설계의 중요한 원칙으로는 시나리오 지향성, 간결성, 그리고 URI서버 지향성을 들 수 있다. 시나리오 지향성은, 구축된 온톨로지에 기반한 최종 서비스 관점에서 불필요한 요소들은 온톨로지 구성요소에 포함시키지 않는다는 것을 의미한다. 온톨로지의 간결성은 규칙에 의해 기술될 수 있는, 유도 가능한 객체관계속성(object property)들을 온톨로지 스키마에 포함시키지 않는다는 것을 의미한다. URI서버 지향성은, 클래스 인스턴스의 고유성 보장을 위해 개별 클래스에 종속적인 URI할당지침에 따라 인스턴스 ID를 부여한다는 것과, 클래스에 의존적인 데이터타입속성(datatype property)들은 온톨로지 스키마에 표현하지 않고 URI서버에서 관리하도록 한다는 것을 의미한다. 위의 후자의 특성은 전술한 온톨로지의 간결성 원칙과도 부합되는 것이다.

온톨로지 기술 언어로는 W3C의 시멘틱웹 표준 언어인 OWL을 사용하였고, 온톨로지 편집도구로는 Protégé 3.1.1을 이용하였다.

2 온톨로지 스키마

과학기술 연구/개발 분야를 다루는 *기반정보 온톨로지*는 현재 핵심적인 최상위 클래스로 **인력**, **기관**, **과제**, **저작물**, **게재지**, **토픽**을 포함하고 있다. **저작물**은 **논문**, **특허**, **연구보고서**의 하위클래스들로 세분되며, **게재지**는 **학술지**와 **학술대회논문집**의 하위클래스를 갖는다.

위의 핵심 상위 클래스들 간의 최소한의 객체관계속성(object property)을 나열하면 다음과 같다.

저작물 - hasCreatorsInformation (저작자정보를 갖는다) - 저작자정보
저작물 - hasOriginatedProject (유발과제를 갖는다) - 과제
저작물 - hasPublication (게재지를 갖는다) - 게재지
저작물 - hasTopic (토픽을 갖는다) - 토픽
저작자정보 - hasCreator (저작자를 갖는다) - 인력
저작자정보 - hasOrganizationOfCreator (저작당시기관을 가진다) - 기관
과제 - hasOrganizationOfFundingProject (발주기관을 가진다) - 기관
과제 - hasOrganizationOfPerformingProject (수탁기관을 가진다) - 기관
토픽 - hasSubTopic (서브토픽을 가진다) - 토픽
인력 - hasOrganizationOfPerson (현재소속기관을 가진다) - 기관

위에서, **저작자정보**라는 핵심 상위 클래스가 있는데, 이것은 저작물의 특정 순위 저자를 모델링한 것으로, 저작물 작성 당시의 저자가 갖는 속성들을 포함한다. **저작자정보**는 데이터타입속성(datatype property)으로 **저자순위**, **저자기여도**를 가지며, **저작물**, **인력**, **기관** 등의 클래스와 객체관계속성을 맺고 있다. 즉, 한 **저작물**의 특정 순위 저자에 해당하는 **인력**을 표현하기 위해 **저작자정보**와 **인력** 간에 객체관계를 설정하고, 특정 순위 저자의 저작물작성당시의 소속 **기관**을 표현하기 위해 **저작자정보**와 **기관** 간에 객체관계를 설정한 것이다.

기반정보 온톨로지 스키마에서 특별히 주목할 점은 그것의 개별 클래스들이 데이터타입속성을 거의 갖지 않는다는 것이다. 예를 들어, *기반정보 온톨로지*에서 **인력** 클래스는 인력의 이름과 같은 기본적인 데이터타입속성을 갖지 않으며, **과제** 클래스의 경우도 과제의 명칭과 같은 데이터타입속성이 표현되어 있지 않다. 이러한 데이터타입속성들은 특정 클래스에 종속적인 것들인데, *기반정보 온톨로지*에서는 이러한 데이터타입속성들을 URI서버에서 관리하는 구조로 설계되었다. URI서버와 온톨로지 사이는 URI값을 매개로 연결된다. 이것의 구체적인 예제는 4.1절에서 기술된다.

*기반정보 온톨로지*는, 연구개발 분야에서 기존에 개발된 대표적인 온톨로지들인 KA2 온톨로지, AKT 온톨로지, SWRC 온톨로지들과 비교했을 때, 온톨로지 모델링 관점에서의 두드러진 차이는 없다. 하지만, *기반정보 온톨로지*는, 인력이나 기관의 연구성과를 정확히 산정하기 위해 기존 온톨로지에서는 찾아볼 수 없는 **저작자정보**라는 클래스를 도입한 점과 인스턴

1) http://www.aktors.org/publications/ontology/
2) http://protege.stanford.edu/ontologies/ontologyOfScience/ontology_of_science.htm

스의 고유한 표현을 위해 동명이인의 중의성을 해소하는 전처리 과정을 거친다는 점이 독특하다. 동명이인의 중의성 해소에 대해서는 4.2절에서 기술할 것이다.

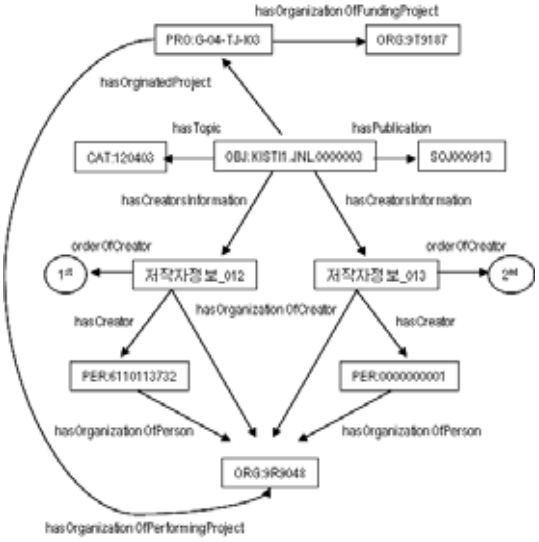
IV. URI지향적 온톨로지 인스턴스 표현

1. 기반정보 온톨로지 인스턴스 표현의 예

그림 1은 *기반정보 온톨로지*의 스키마에 대응하는 인스턴스 표현의 한 예를 보여준다. 그림 1에 대응하는 실세계 정보는 아래와 같다.

[과제]
- 과제명: 생물 유전자원정보 네트워크 구축 및 관련기술연구
- 과제기간: 2004~2005
- 발주/수탁기관: 과학기술부/한국과학기술정보연구원

[과제의 유발성과물]
- 논문제목: 생물다양성데이터 활용을 위한 국가 데이터노드 구축
- 저자: 박형선, 안성수, ...
- 게재지명: 지식정보인프라
- 논문토픽: 유전자, ...



URI	유형	메타데이터
PER:6110113732	인력	이름: 박형선, 소속기관: ORG:9R9048
PER:0000000001	인력	이름: 안성수, 소속기관: ORG:9R9048
OBJ:KISTI1.JNL.0000003	저작물(논문)	제목:생물다양성데이터 활용을 위한 국가 데이터노드 구축 관련과제: PRO:G-04-TJ-K03
PRO:G-04-TJ-K03	과제	제목: 생물 유전자원정보 네트워크 구축 및 관련기술 연구 연도: 2004-2005, 발주기관:ORG:9T9187, 수탁기관: ORG:9R9048
ORG:9T9187	기관	기관명: 과학기술부
ORG:9R9048	기관	기관명: 한국과학기술정보연구원
PUB:SOJ000913	게재지	발행기관: ORG:9R9048, 게재지명: 지식정보인프라
CAT:120403	토픽	용어명: 유전자

▶▶그림 1. URI 지향적 기반정보 온톨로지 인스턴스 표현의 예 (표의 내용은 URI서버에 저장된 데이터이며, URI값의 생략된 namespace는 http://www.kisti.re.kr/isrl/NSTRDOnto.owl#이다)

그림 1에서 개별 인스턴스들은 그것의 고유한 URI값으로 표현되고 있다. *기반정보 온톨로지*에서 각 클래스들은 그것의 인스턴스 ID를 할당하는 URI할당지침이 관련되어 있다. 예를 들어, **저작물** 인스턴스는 KOI[3]식별체계를 간략화한 지침에 따라 URI값이 부여되고, 인력 인스턴스에는 국가과학기술인력종합정보시스템[4]에서 사용하는 인력식별체계를 근간으로 하는 URI값이 할당된다. 그림 1에서 알 수 있듯이 클래스 종속적인 데이터타입속성값들은 거의 온톨로지에 표현되지 않으며, URI서버에서 관리된다. 온톨로지와 URI서버와의 연결은 개별 인스턴스의 URI값으로 가능하다.

2. 인스턴스의 고유성 확보 방안

*기반정보 온톨로지*의 개별 인스턴스에 고유한 URI 값을 대응시키기 위해서는 4.1절에 소개한 클래스별 URI식별체계를 적용하기 이전에 학술정보 데이터에 내재된 인스턴스에 대한 언어적 표현의 동의성과 다의성 문제를 해소할 필요가 있다. 위의 두 문제는 실세계의 인스턴스가 고유한 URI값 대신 동의성과 다의성을 내재하고 있는 자연언어로 표현된다는 데서 기인한다. 예를 들면, 학술데이터에서 **인력** 인스턴스는 이름으로 표현되며, **기관** 인스턴스는 고유한 기관코드값 대신 기관명으로 표현된다.

동일 인스턴스에 대한 서로 다른 표현의 문제인 동의성은 *기반정보 온톨로지*의 경우 그리 심각하지는 않다. 그 이유는 그러한 동의성의 문제가, 그들의 동의어 표현의 집합이 어느 정도 닫혀 있다고 볼 수 있는 기관명이나 게재지명에 국한되기 때문이다. 그러나, **인력**의 이름이 갖고 있는, 서로 다른 인스턴스에 대한 형태적 동일 표현의 문제인 다의성은 실세계의 신원이 온톨로지 내에서 왜곡돼는 결과를 초래하므로 반드시 해소되어야 한다. 본 연구팀에서는 이러한 동명다인의 문제를 해소하기 위한 단서로 동일명 연구자의 논문게재건들에 내재된 공저자정보, 전자메일주소, 소속기관명, 출판년도 등을 복합적으로 고려하는 방법을 개발하였다. 이 논문에서는 그 방법의 핵심 아이디어만을 소개한다. 먼저 아래와 같은 동일명 연구자 A의 논문게재건들이 주어졌다고 하자.

논문 1의 저자명: A, B, C
논문 2의 저자명: A, C, D

위에서 논문 1과 2는 A의 공동저자 C를 공유하고 있는데, 이러한 상황에서 논문 1과 2에서의 두 A는 동명이인일 확률이 매우 낮을 것이다. 여기에 더하여 만약 논문 1과 2의 두 A의

3) KOI(Knowledge Object Identifier): 한국과학기술정보연구원에서 개발한 과학기술지식정보를 연계/공유하기 위한 표준식별체계.
4) http://www.hrst.or.kr/

전자메일 주소가 같고, 논문 1과 2의 두 C의 전자메일 주소도 같다면 논문 1과 2의 두 A가 동일인일 확률은 더 높아진다. 여기에 논문 1과 2의 두 A와 두 C가 각각 소속기관까지 같고, 두 논문의 출판년도마저도 같다면, 논문 1과 2의 A라는 동일 이름의 두 저자는 실제로 같은 사람일 확률이 훨씬 더 높아질 것이다. 이러한 동명다인 해소 방법을 컴퓨터로 자동화시키는 것은 어렵지 않을 것이다.

V. 결 론

논문에서는 연구개발 분야의 온톨로지의 한 예로써 *기반정보 온톨로지*에 대해 기술하였다. 현재 *기반정보 온톨로지*의 인스턴스 DB는 한국과학기술정보연구원의 내부 성과정보와 최근 국내 학술대회논문을 중심으로 구축되고 있으며, 학술지논문을 포함시킬 예정에 있다. 향후 시멘틱웹 분야의 온톨로지들은 재사용 장려 측면에서 콘텐츠 관점으로 접근하여 온톨로지의 수집/검증/저장/검색을 담당하는 기구의 마련이 필요할 것이다.

▌참 고 문 헌▌

[1] Benjamins, V. R., Fensel, D., Decker, S., and Gomez-Perez, A.. (1999) "(KA)2: Building ontologies for the internet: a mid term report", International Journal of Human -Computer Studies, 51(3), pp.687-712.

[2] Gruber, T. R. (1993) "A translation approach to portable ontology specifications", Knowledge Acquisition, 5(2), pp.199-220.

[3] Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., Oberle, D. (2005) "The SWRC Ontology - Semantic Web for Research Communities", In Proceedings of the 12th Portuguese Conference on Artificial Intelligence, pp.218-231.