시맨틱 웹 서비스를 위한 온톨로지 기반 주제어 공통 표현 모델

정한민O 김평 이미경 성원경
정보서비스연구팀, 한국과학기술정보연구원
jhm@kisti.re.kr

Topic Keyword Common Representation Model Based on Ontology for Semantic Web Services

Hanmin JungO Pyung Kim MI-Kyoung Lee Won-Kyung Sung
ISRL., KISTI

요 약

주제어는 정보 서비스를 비롯한 여러 응용 분야들에서 유용하게 사용되는 지식이지만, 주제어 간 관계가 다양할 뿐만 아니라 목적에 맞도록 개별적으로 설계됨으로써 주제어 관계 속성 유형과 무관하게 공유가 가능한 주제어 공통 표현 모델이 제시되지 못하였다. 본 연구는 응용 분야, 온톨로지 종류와 무관하게 적용될 수 있으며 시맨틱 웹 서비스 간 공유가 가능한 주제어 공통 표현 모델을 제시하고자 한다. 이를 위해, 주제어 관계를 범용 클래스로 정의하고, 주제어 관계 속성 유형을 데이터타입 속성 (Datatype Property)으로 선언하였다. 또한, 주제어 역시 그 속성 유형을 데이터타입 속성으로 선언하였는데, 결국 다양한 유형의 관계들을 용이하게 표현할 수 있도록 하기 위한 것이다. 실험을 위해 주제어 간 관계수가 70,804,233개이며 주제어 관계 속성 유형이 4가지인 과학 기술 기반 정보 온톨로지와 주제어 간 관계수가 44,147개이며 주제어 관계 속성 유형이 13가지인 표준 정보 온톨로지를 대상으로 본 연구에서 제안한 주제어 공통 표현 모델을 적용하였으며 총 284,744,802 개의 RDF (Resource Description Framework) Triple을 생성하였다.

1. 서론

서로 다른 응용 분야들에의 시맨틱 웹 서비스를 구현하기 위해서는 다중 온톨로지가 필요하다. 그렇지만, 다중 온톨로지가 반드시 모든 클래스들을 상호 베타적으로 정의해야 하는 것은 아니다. 예를 들어, 'Person' 클래스의 경우에는 인력을 다루는 모든 응용 분야들에서 공통적으로 사용할 수 있다. 보편적으로 인정되는 'Person' 클래스의 속성들이 정의될 수 있기 때문이다. 주제어의 경우에는 상황이 다른데, 그 이유는 그 쓰임의 다양성에 있다. 기계 번역에 사용하는 경우에는 대역 정보 (예. 영어, 일본어 등) 중심으로 구축될 것이며, 용어 브라우징에 사용하는 경우에는 다양한 용어 간 관계가 구축될 것이며, 시소러스를 구축하는 경우에는 상·하위 관계 중심의 용어 간 관계 가 구축될 것이다. 이러한 응용 분야들이 서로 독립적으로 구축되는 경우에는 온톨로지 간 공유를 고려하지 않을 수 있지만, 본 연구와 같이 동시 다발적으로 여러 응용 분야들을 구축하는 경우에 있어서는 효율성과 재활용성이 중요한 이슈가 된다. 특히, 사전, 시소러스 등의 언어 자원과 같이 막대한 구축 비용이 드는 콘텐츠에 대해서는 더욱 중요한 이슈가 될 수밖에 없다.

본 연구에서 공유 대상으로 선정한 것은 주제어이다. 학술 연구 정보 분석 서비스, 표준 정보 서비스를 포함한 공공재적 성격의 응용 분야에서는 주제어 중심의 서비스가 선호된다. [1]에서 언급한 바와 같이 이러한 분야에서의 사용자 질의는 분야명이나 영역 내 주제어를 포함하는 경향이 많기 때문이다. 그렇지만, 대부분의 연구들은 해당 응용 분야에 맞게 주제어를 모델링하고 구축한 것인데, 그 이유는 구축 비용이 크게 발생하는 주제어 자원을 다양하게 확보하는 것이 현실적으로 어렵기 때문일 것이다. 물론, 최근 관심의 대상이 되는 온톨로지 간 매핑이 하나의 대안이 될 수는 있으나, 매핑되는 온톨로지 내의 개념이나 관계들이 정확히 일치한다는 보장이 없고, 매핑의 정확도가 항상 이슈로서 제기되므로 근본적인 해결책으로 보기 어렵다. 이에 본 연구는 시맨틱 웹 프레임워크를 다양한 응용 분야에 적용하면서 확보한 주제어 자원을 시맨틱 웹 프레임워크 상에서 공유하고 재활용하기 위한 목적을 가지고 온톨로지 기반 주제어 공통 표현 모델을 구축하고자 한다 [2]. 특히, 다중 온톨로지에서 주제어 표현 모델을 공유함으로써 다른 시맨틱 웹 프레임워크로의 이식성을 보장할 수 있도록 하는데 초점을 맞춘다. 또한, 추론 작업이 가능하도록 표준 온톨로지 언어인 OWL (Web Ontology Language) 기반으로 주제어 공통 표현 모델을 제시하여, 추론 기반 시맨틱 웹 서비스에 바로 적용할 수 있도록 한다.

2장에서는 주제어 표현에 대한 기존 연구들을 살펴보

고, 3장에서는 본 연구의 배경인 시맨틱 웹 프레임워크 상에서의 주제어 공통 표현 모델의 필요성을 언급한다. 4장의 주제어 공통 표현 모델 기술을 통해 이식성과 재활용성 확보 방안을 보이며, 5장의 실험 및 결론을 통해서 본 연구의 의미를 설명하고자 한다.

2. 관련 연구

말뭉치가 급격히 늘어남에 따라 주제어 간 관계를 자동 추출하는 연구들이 많아지고 있으며 이들이 추출한 관계들을 공통적으로 모델링할 수 있는 방안이 필요하다. 본 연구는 다양한 관계를 가진 주제어들을 공통적으로 표현하기 위한 주제어 모델을 제시하는 데 초점을 맞춘다. [3]은 주제어 표현 정보를 이용하여 Is-a 관계를 추출하고 정의문 정보를 통해 그 관계를 검수하는 방법을 제안했으며, [4]는 표준화된 정의 내용을 통해 상위어 (BT; Broader Term), 하위어 (NT; Narrower Term), 관련어 (RT; Related Term)를 추출하여 시소러스를 확장하는 방법을 제안하였으며, [5]는 조어 정보를 이용하여 상·하위 관계, 관련어 관계, 형제 관계를 추출하는 방법을 제안하였다. 4장에서 제시하는 주제어 모델에 의하면 이들 모두 'TopicRelation' 클래스의 'kindOfTopicRelation'에 추출된 관계를 정의하는 방식으로 처리가 가능하다.

[6]의 연구에서는 시소러스 구조를 토픽맵의 요소로 표현하고 시소러스 관계들을 모델링하기 위해 구조 체계 매핑과 구조 체계 병합이라는 복잡한 방법들을 사용하는데, 시소러스에 한정되어 관리할 수 있는 편의성이 제공될 수 있지만, 시소러스 이외의 관계들까지 포용하기에는 범용성이 부족하다. 온톨로지 기반으로 주제 정보 관리를 시도한 또 다른 시도로는 [7]이 있는데, 시민 단체를 대상으로 선정하여 활동 주제 및 내역을 온톨로지 모형으로 표현하고 있다. 그렇지만, 주제를 클래스 수준에서 기술함으로써 클래스 수의 통제를 어렵게 하고, 단지 상·하위 관계만을 모델링했다는 한계를 가진다.

ISO/IEC AWI 19763-04[1]에서 다루는 메타모델 상호운용성을 위한 프레임워크 (Framework for Metamodel Interoperability)에서도 개발의 일관성 있는 모델에 적용하기 위한 목적으로 메타모델 정보 처리 상호운용성과 메타 모델 프레임워크 구조 등을 다루나, 본 연구의 범위에 해당하는 온톨로지 스키마나 모델링 구조 등은 범위에서 제외하고 있다.

기존 표준 시소러스에서 표현하지 못했던 다양한 관점을 반영하고 시소러스 구조에 대한 명시적 설명을 가능하도록 한 패싯 기반 시소러스 연구들도 있었는데 [8] [9] [10], [9]의 경우에는 본 주제어 모델에서 'kindOfTopic' 속성으로 대응될 수 있는 개념 패싯과 'kindOfTopicRelation' 속성으로 대응될 수 있는 관계 패싯을 도입하여 개념 간의 관점을 기술하였다 (그림 4 참조). 또한, 시소러스의 개념 관계를 확장하여 2차 개념 관계, 본 연구에서의 주제어 관계 속성 유형과 유사한, 를 제안한 연구도 최근에 소개되었다 [11]. [9]의 연구와 같이 개념 관계의 관점을 세분하여 기술한다는 특징을 가지지만 본 연구에서 다양한 개념 관계를 'TopicRelation' 클래스와 'kindOfTopicRelation' 속성으로 단순화시켜 모델링한데 반해 UF, RT, TIT 등 1차 개념 관계를 유지하면서 2차 개념 관계를 추가함으로써 개념 관계 추가의 유연성이 떨어진다는 단점이 있다.

용어 관계를 체계적으로 분석하고자 하는 시도를 한 [12]는 용어 관계 구성을 동일성 (예. USE-UF), 계층성 (예. BT-NT), 연관성 (예. RT)으로 나누어 살펴보았다. 그렇지만, 상기와 같은 전형적인 구분 조차 명확하지 않으며 용어 관계 유형의 수와 다양성이 크다는 점을 결론에서 지적하고 있다. 또한, 용어 관계 체계화의 기본적인 기준은 용어의 의미가 되어야 한다고 주장하는데, 이는 개념과 개념을 표현하는 용어 간의 관계를 잘 모델링할 필요가 있음을 의미한다고 볼 수 있다. 결국, URI (Uniform Resource Identifier), RDF (Resource Description Framework), OWL (Web Ontology Language)과 같은 체계적인 시맨틱 웹 표준을 활용하여 간단명료한 방식으로 개념과 용어를 모델링함으로써 관리적 편의성과 지식 체계 확장의 용이성을 확보할 수 있을 것이다.

3. 시맨틱 웹 프레임워크와 다중 온톨로지

시맨틱 웹 프레임워크는 플랫폼, 응용 분야 독립적으로 시맨틱 웹 서비스를 지원할 수 있도록 개발되어야 한다. 그렇지만, 기반 지식이 되는 온톨로지는 응용 분야에 의존적으로 설계되고 구축되는데, 이때 온톨로지 재활용이 효율성을 향상시키는 중요한 요소가 된다. 특히, 정보 서비스 분야에서의 주제어는 정보 검색, 질의응답 등 핵심 서비스들을 이용하는 사용자들에게 반드시 제공되어야하는 정보이다. 주제어는 특정 응용 분야의 검색 질의어 중 많은 부분을 차지하는데 [1], 단순히 검색 질의어로서의 역할 뿐만 아니라 주제어 간 관계를 그림 1과 같이 브라우징 방식으로 사용자에게 제시함으로써 서비스의 연속성을 높여줄 수 있다.

1) http://kisi.or.kr/information/project_detail.asp?page=1&idx=947&selCd=&txtKey=mapping

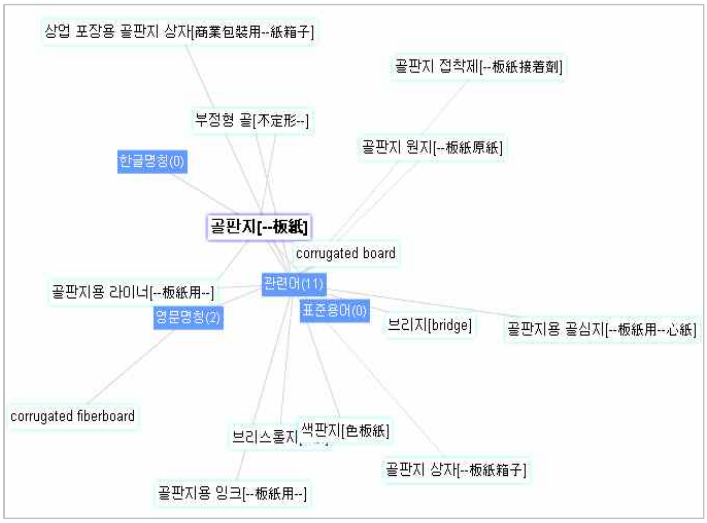

그림 1. 주제어 브라우징 예 (주제어 관계는 '한글명칭', '관련어', '표준용어', '영문명칭'이며 나머지 노드들은 주제어들임)

그림 2와 3은 시맨틱 웹 프레임워크 OntoFrame 상에서 모델링된 온톨로지들이다 [2]. 그림 2를 보면, 주제어 관계 속성 유형을 개체 속성 (Object Property)로 선언함으로써 새로운 주제어 관계 속성 유형을 추가하거나 외부에서 구축된 주제어 및 주제어 관계를 매칭시키는 것을 어렵게 한다. 반면에, 그림 3에서와 같이 주제어 관계를 범용 클래스로 정의하고, 주제어 관계 속성 유형을 데이터타입 속성 (Datatype Property)으로 선언하면, 다중 온톨로지에서도 주제어 표현 모델은 유연한 공유를 가능하게 할 수 있다. 즉, 서로 다른 환경에서 구축된 주제어 정보들을 통합하여 관리하고 서비스할 수 있도록 해줄 수 있는 것이다.

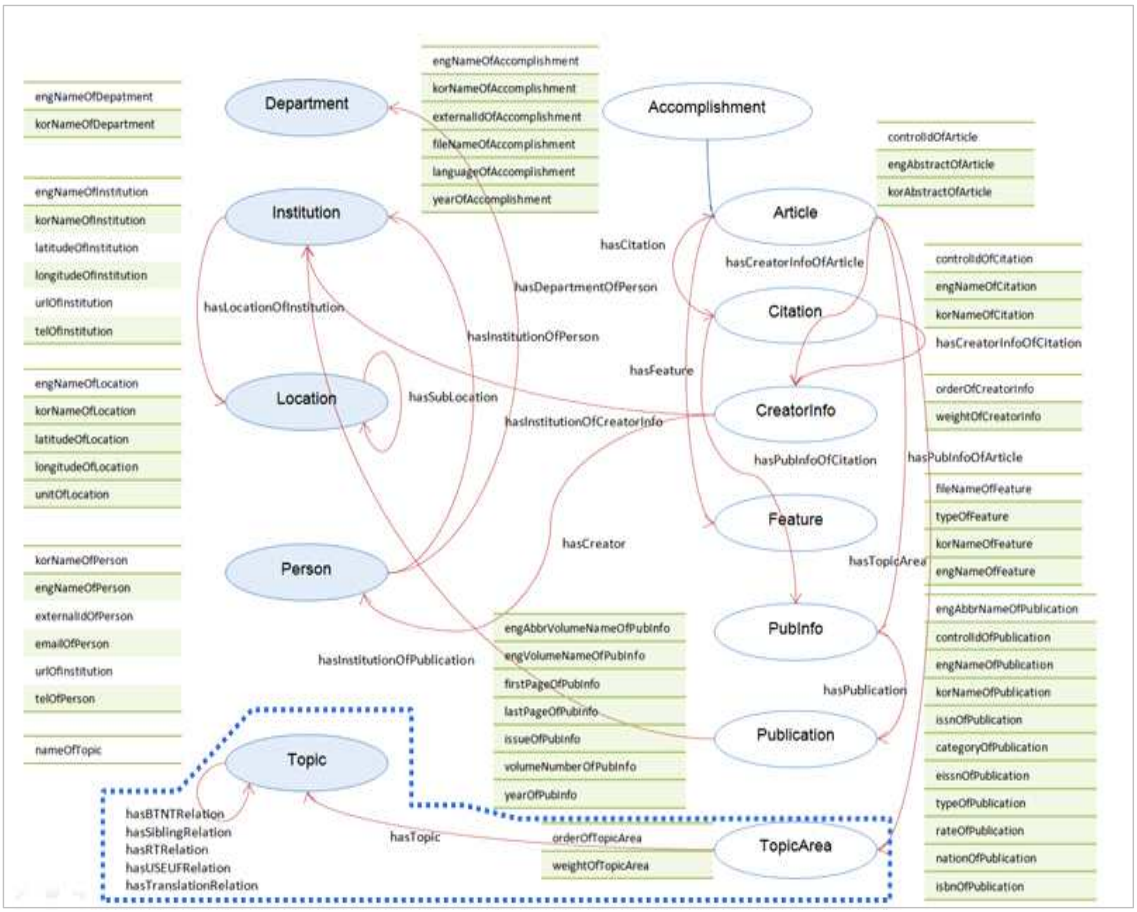

그림 2. 시맨틱 웹 서비스에 특화된 주제어 표현 모델이 적용된 온톨로지 예 (주제어 관계 속성 유형을 각각 개체 속성으로 표현한 모델, 점선 영역 내 참조)

4. 주제어 공통 표현 모델

용어는 특정 주제에서 사용되는 하나의 개념을 유일하게 나타내는 주요한 단위이다 [4]. 본 논문에서 다루는 주제어는 전문 용어에 해당하나, 그림 4와 같이 개념 표현을 위한 'Topic' 클래스 내에 포함될 수 있게 함으로써

개념 수준에서 용어를 다룰 수 있도록 한다. 즉, 동형이의어에 대해 서로 다른 'Topic' 인스턴스를 갖도록 함으로써 개념적 분리를 가능하게 한다는 것이다.

그림 4는 다중 온톨로지 상에서 공통으로 사용할 수 있는 주제어 온톨로지 스키마를 보여준다. 온톨로지 내 클래스들은 'hasTopic' 개체 속성 (Object Property)으로 주제어 ('Topic') 클래스와 연결된다. 주제어 관계 속성 ('TopicRelation') 클래스는 주제어 클래스들을 연결하는데 사용되는데, 주제어 클래스는 주제어 관계 속성 클래스와 'hasTopicRelation'으로, 주제어 관계 속성 클래스는 주제어 클래스와 'hasTopicOfTopicRelation'으로 연결된다.

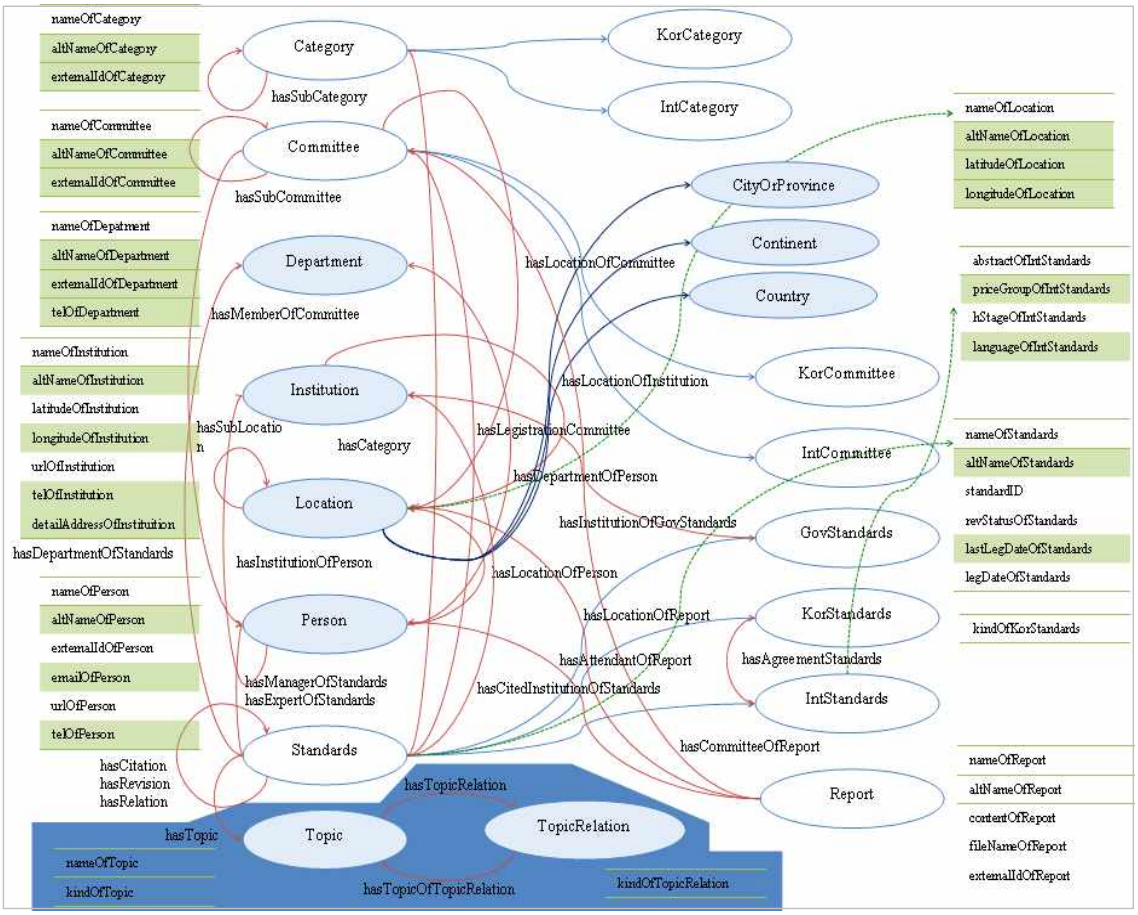

그림 3. 주제어 공통 표현 모델이 적용된 온톨로지 예 (주제어 관계 속성 유형을 주제어 관계 클래스 내 데이터타입 속성으로 표현한 모델, 색칠된 영역 내 참조)

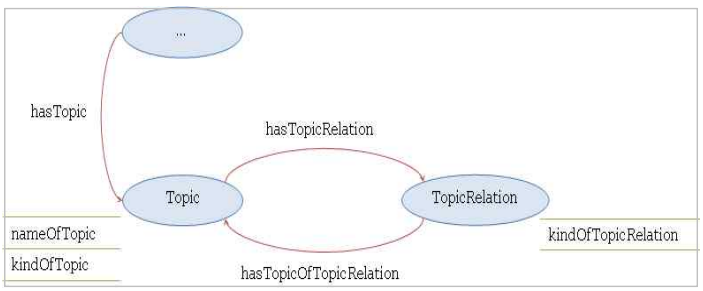

그림 4. 주제어 공통 표현 모델 (타원은 클래스, 클래스 간 화살표는 개체 속성 (Object Property), 클래스 옆 테이블은 데이터타입 속성 (Datatype Property))

주제어 속성에는 주제어 명칭을 의미하는 'nameOfTopic'과 주제어 속성 유형을 의미하는 'kindOfTopic'이 있으며, 주제어 관계 속성에는 주제어 관계 속성 유형을 의미하는 'kindOfTopicRelation'이 있다. 이와 같이 주제어 속성, 주제어 관계 속성을 데이터타입 속성으로 정의한 것은 다양한 유형의 관계를 용이하게 표현할 수 있도록 하기 위한 것이다. 이들을 개체 속성으로 정의한다면, 새로운 유형의 관계에 따른 온톨로지 스키마의 확장이 불가피할 수밖에 없기 때문이다. 그림 4를 OWL (Web Ontology Language)로 기술하면 다음과 같다. 주제어 인스턴스에서 주제어 명칭이 하나만 존재

해야 하므로 이 속성은 함수 속성 (Functional Property)으로 정의되지만, 주제어 속성 유형과 주제어 관계 속성 유형은 이러한 제약을 두지 않는다.

```
<owl:Class rdf:ID="TopicRelation"/>
<owl:Class rdf:ID="Topic"/>
<owl:ObjectProperty rdf:ID="hasTopicRelation">
 <rdfs:range rdf:resource="#TopicRelation"/>
 <rdfs:domain rdf:resource="#Topic"/>
</owl:ObjectProperty
<owl:ObjectProperty rdf:ID="hasTopicOfTopicRelation">
 <rdfs:range rdf:resource="#Topic"/>
 <rdfs:domain rdf:resource="#TopicRelation"/>
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:ID="kindOfTopicRelation">
 <rdfs:range
 rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
 <rdfs:domain rdf:resource="#TopicRelation"/>
</owl:DatatypeProperty>
<owl:FunctionalProperty[2] rdf:ID="nameOfTopic">
 <rdf:type
 rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
 <rdfs:domain rdf:resource="#Topic"/>
 <rdfs:range
 rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:FunctionalProperty>
<owl:DatatypeProperty rdf:ID="kindOfTopic">
 <rdfs:range
 rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
 <rdfs:domain rdf:resource="#Topic"/>
</owl:DatatypeProperty>
```

'Topic' 클래스의 'kindOfTopic'은 주제어의 속성 유형을 지칭하는데, 시맨틱 웹 서비스에서 속성 유형에 따른 차별화된 정보를 제공하는데 활용된다. 예를 들어, 주제어 "Prion"의 'kindOfTopic'이 "Protein"인 경우에 검색 결과에서 해당 단백질의 구조들을 추가로 보여줄 수 있다 (그림 5 참조).

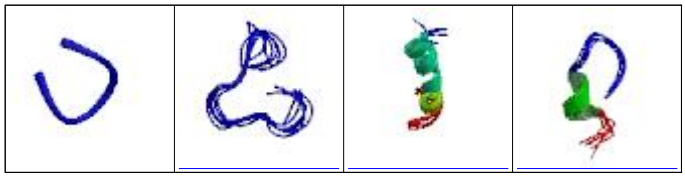

그림 5. 검색 결과 내 단백질 구조 예 ("Prion" 단백질 내 구조들)

'kindOfTopic'이 "Gene"인 경우에는 GenBank[3] 내 유전자 정보를 추가로 보여줄 수 있듯이, 주제어의 속성 유형은 서비스 인터페이스 구성에 영향을 미친다.

5. 실험 및 결론

본 연구에서는 과학 기술 기반 정보 온톨로지 (그림 2 참조; 본 연구에서 제시한 주제어 공통 표현 모델에 맞도록 수정된 버전은 그림 3을 참조하면 되므로 본 논문에서는 생략함)와 표준 정보 온톨로지 (그림 3 참조)를 대상으로 주제어를 공유할 수 있도록 모델링을 수행하였다.

과학 기술 기반 정보 온톨로지의 경우에는 전체 주제어 수가 652,507개이며, 주제어 간 관계수는 70,804,233개이다. 주제어 관계 속성 유형 'kindOfTopicRelation'은 "BT (상위어)", "NT (하위어)", "Sibling (형제어)", "RT (관련어)"를 그 값으로 가진다. 이를 RDF Triple로 기술한 예를 살펴보면 다음과 같은데, "TOP_"는 주제어클래스의 인스턴스를, "REL_"는 주제어 관계 속성 클래스의 인스턴스를 의미한다. 전체 Triple 수는 284,521,946개이다.

```
:TOP_515 ref:hasTopicRelation :REL_9721993 .
:REL_9721993 ref:hasTopicOfTopicRelation :TOP_1634 .
:REL_9721993 ref:kindOfTopicRelation "Sibling"^^xsd:string .
```

표준 정보 온톨로지의 경우에는 전체 주제어수가 23,134개이며, 주제어 간 관계수는 44,147개이다. 그렇지만 주제어 관계 속성 유형은 보다 다양하여 "GER (독일어)", "KEN (영어-한국어)", "UF (비우선어)", "FRA (불어)", "ENG (영어)", "KGE (독일어-한국어)", "RT (관련어)", "KFE (불어)", "KAL (알바니아어-한국어)", "ALB (알바니아어)", "USE (우선어)", "BT (상위어)", "NT (하위어)"를 그 값으로 가진다. 이를 RDF Triple로 기술하면 다음과 같으며, 전체 Triple 수는 222,856개이다.

```
:TOP_18571 ref:nameOfTopic "기준 래크 피치선 위치[基準--線位置]"^^xsd:string .
:TOP_7220 ref:nameOfTopic "격자 거더[格子--]"^^xsd:string .
:TOP_5820 ref:nameOfTopic "갑판보 브래킷[甲板--]"^^xsd:string .
.....
:TOP_945 ref:hasTopicRelation :TRE_6593 .
:TRE_6593 ref:hasTopicOfTopicRelation :TOP_948 .
```

2) 각 인스턴스에 대해 하나의 값만을 가질 수 있는 속성

3) http://www.ncbi.nlm.nih.gov/Genbank/

:TRE_6593 ref:kindOfTopicRelation "ENG"^^xsd:string .

이와 같이 주제어간 다양한 관계를 가지는 경우에도 본 연구에서 제안한 주제어 공통 표현 모델을 이용하여 간단하게 표현할 수 있다. 또한, URI (Uniform Resource Identifier)로 개념화를 시키고 인스턴스로서 주제어를 부착하는 방식으로 온톨로지를 표현함으로써 동일 주제어가 서로 다른 의미를 가지는 동형이의어에 대해서도 명확하게 구분할 수 있다 (그림 6 참조).

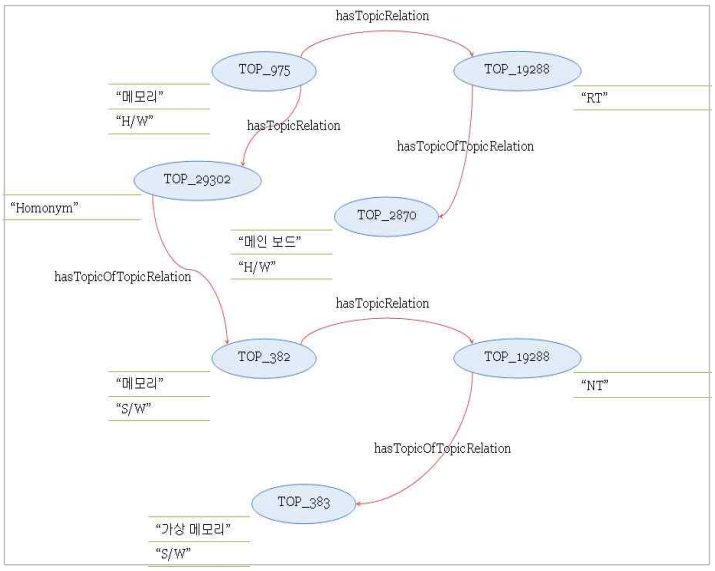

그림 6. 동형이의어를 가진 주제어 공통 표현 모델 예 (“메모리”라는 주제어들에 대해 동형이의어가 존재하며, 이 둘은 "TOP_29302"에 의해 연결되어 있으며, 각각은 "TOP_2870", "TOP_383"을 통해 또 다른 주제어와 연결되어 있다.)

주제어는 정보 서비스를 비롯한 여러 응용 분야들에서 유용하게 사용되는 지식이지만, 주제어 간 관계가 다양할 뿐만 아니라 목적에 맞도록 개별적으로 설계됨으로써 범용적으로 적용할 수 있는 주제어 공통 표현 모델이 제시되지 못하였다. 본 연구는 응용 분야, 온톨로지에 관계없이 독립적으로 운용될 수 있고 공유가 가능한 방식으로 주제어 공통 표현 모델을 제시하였다는데 그 의미가 있다. 향후 연구를 통해서 공개 가능한 수준의 주제어들을 인스턴스로 탑재한 주제어 온톨로지를 독립적으로 구축하고 서비스할 수 있을 것으로 기대한다. 여기에는 [13]의 연구에서 제시된 35개의 다양한 의미적 관계를 포함시킬 수도 있을 것이다. 또한, PDB (Protein DB)[4], GenBank 등으로부터 주제어에 상당하는 전문 용어들을 획득하여 본 연구에서 제안한 주제어 공통 표현 모델에 반입하는 계획도 가지고 있다.

4) http://www.rcsb.org/pdb/home/home.do

참고문헌

[1] G. McCrakken, "Vertical Search - Challenges and Opportunities", AOP Vertical Search (Presentation Material), 2008.

[2] W. Sung, H. Jung, P. Kim, I. Kang, S. Lee, M. Lee, D. Park, and S. Hahn, "A Semantic Portal for Researchers Using OntoFrame", In Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, 2007.

[3] 황금하, 이신목, 남윤영, 신지애, 최기선, “시소러스를 이용한 온톨로지의 Is-a 관계 설정”, 한국정보과학회 추계학술대회, 2006.

[4] 김지훈, 김태수, “용어정의와 관계추출을 통한 시소러스 확장에 관한 연구”, 한국문헌정보학회지 40(1), 2006.

[5] 정한민, 이미경, 성원경, “조어중심적 주제어간 관계 추출 및 분석”, 한국언어정보학회 정기학술대회, 2008.

[6] 남영준, “토픽맵을 이용한 시소러스의 구조화 연구”, 정보관리학회지 22(3), 2005.

[7] 유사라, “특정주제 정보관리를 위한 온톨로지 모형 연구”, 정보관리학회지 24(1), 2007.

[8] 정영미, 김명옥, 이재윤, 한승희, 유재복, “과학기술 분야 통합 개념체계의 구축 방안 연구, 한국정보관리학회지 19(1), 2002.

[9] 황순희, 정한민, 성원경, “패싯(Facet)을 이용한 과학기술분야 시소러스 구축과 활용방안”, 정보관리연구 37(3), 2006.

[10] 김원중, 양재동, “시각화된 환경에서 다차원 관점을 지원하는 객체기반 패싯 시소러스 관리 시스템 모델의 정형화 및 구현”, 정보과학회논문지: 소프트웨어 및 응용 34(9), 2007.

[11] 이정연, 김정민, 최석두, 김이겸, “기초학문자료 메타데이터 설계 분석 및 온톨로지 적용 방안 연구”, 한국문헌정보학회지 41(2), 2007.

[12] 백지원, 정연경, “지식조직체계의 용어관계 유형에 관한 연구”, 한국문헌정보학회지 39(4), 2005.

[13] D. Moldovan, A. Badulescu, M. Tatu, D. Antohe, and R. Girju, "Models for the Semantic Classification of Noun Phrases", In Proceedings of the Workshop on Computational Lexical Semantics in HLT-NAACL, 2004.