# URI 서버에 기반한 국가 R&D 기반정보 온톨로지 설계 및 구현

# A Design and Implementation of National R&D Reference Information Ontology Based on URI Server

정 한 민* · 강 인 수** · 구 희 관*** · 이 승 우**** · 성 원 경*****

Han-min Jung · In-Su Kang · Hee-Kwan Koo · Seung-woo Lee · Won-Kyung Sung

차 례



초 록

시맨틱 웹의 발전은 정보의 규격화, 의미화를 통한 지식을 기본으로 이루어지며, 온톨로지는 이러한 지식표현을 위해 필수적으로 사용되는 도구이다. 온톨로지상에서 개체(Individual)들은 URI(Uniform Resource Identifier)를 이용하여 유일하게 지칭될 수 있어야 한다. 예를 들어, 국가 R&D 기반정보를 모델링하고, 이를 이용하고자 하는 경우에 URI 기반의 온톨로지 설계와 구현이 필수적으로 요구된다. 그렇지만, 식별체계나 URI를 사용하기 위해서는 방대한 인적·물적 자원의 투입이 불가피하여 과학기술문헌상의 인력정보를 식별체계 기반으로 구축하고자 하는 시도가 미약한 실정이었다. 이에 본 연구는 과학기술문헌을 포함한 국가 R&D 기반정보 온톨로지 구축에서 핵심이 되는 인력정보를 포함한 다양한 정보들을 URI 기반으로 구축, 관리, 서비스하는 방법을 기술한다. 약 7,000여건의 국내학술대회 논문들로부터 획득한 기반정보는 추론 서비스를 통해 연구자 네트워크 분석, 성과통계 등 다양한 시맨틱 웹 응용 분야들에 적용된다.

키 워 드

국가 R&D 기반정보 온톨로지, 시맨틱 웹, 추론, URI 서버

* 한국과학기술정보연구원 정보시스템연구팀
(Senior Researcher, Information System Research Lab., KISTI, jhm@kisti.re.kr)
** 교신저자, 한국과학기술정보연구원 정보시스템연구팀
(Senior Researcher, Information System Research Lab., KISTI, dbaisk@kisti.re.kr)
*** 한국과학기술정보연구원 정보시스템연구팀
(Student Researcher, Information System Research Lab., KISTI, hkkoo@kisti.re.kr)
**** 한국과학기술정보연구원 정보시스템연구팀
(Senior Researcher, Information System Research Lab., KISTI, swlee@kisti.re.kr)
***** 한국과학기술정보연구원 정보시스템연구팀
(Principal Researcher, Information System Research Lab., KISTI, wksung@kisti.re.kr)
• 논문접수일자 : 2006년 5월 22일
• 게재확정일자 : 2006년 6월 13일

**ABSTRACT**

The development of Semantic Web basically requires knowledge which is induced by the formalization and semantization of information, and thus ontology should be introduced as a knowledgization tool. URI(Uniform Resource Identifier) is an indispensible scheme to uniquely indicate individuals on ontology. However, it is difficult to find the use cases of identifiers or URIs in real data sets including science & technology publications. This paper describes the method to construct, manage, and serve reference information based on URI which is a crucial component on establishing national R&D reference information ontology. We expect the reference information which was acquired from about 7,000 proceeding papers would be adopted to Semantic Web applications such as researcher network analysis and outcome statistics.

**KEYWORDS**

National R&D Reference Information Ontology, Semantic Web, Inference, URI Server

## 1. 서 론

시맨틱 웹의 발전과 함께 온톨로지 구축의 중요성이 점점 강조되고 있는 현실에서, 온톨로지의 기반이 되는 URI(Uniform Resource Identifier)가 필수적으로 요구되고 있다. 그렇지만, 이러한 중요성에도 불구하고 실험 데이터가 아닌 실제 데이터로부터 정보를 추출하여 정보 간 충돌을 해소하고 URI를 부여하고자 하는 시도가 CS AKTiveSpace[1] 등 외국의 몇몇 사례들을 제외하고는 찾아보기 힘든 형편이다. 국가과학기술인력 종합정보시스템[2]에서와 같이 인력정보에 대해 식별자 기반으로 서비스를 제공하는 곳이 있기는 하나, 해당 정보는 실 데이터로부터 가공된 것이 아닌 사용자가 직접 등록한 정보에 불과하고, 이 또한 해당 시스템 내에서만 식별 가능한 URI 기반이 아닌 단순 식별체계에 기반하고 있다. 실 데이터로부터 발생할 수 있는 다양한 문제들을 해결할 수 있는 방안을 제시하지 않고서는 지속적인 정보확장이 어려울 수밖에 없다.

기존에 논문으로부터 저자정보를 포함한 서지정보를 자동추출하여 구축하는 연구가 있었다(장대근 외 1998). 그렇지만, 저자정보의 의

1) http://www.aktors.org/technologies/csaktivespace
2) http://www.hrst.or.kr/hrst/index.jsp

미적 구축이 아닌 문자인식 기반 문자열추출에 초점이 맞추어졌을 뿐이다. 다만, 식별체계구축 시스템 간의 연계에 대한 연구가 이상환 외(2004)와 신동구 외(2005)를 통해 조금씩 이루어지고 있는데, 본 연구에서 식별체계 재활용 측면에서 과학기술문헌에 부여하는 ID를 KOI(Knowledge Object Identifier) 기반으로 구축함으로써 이러한 연구결과를 적극 반영한다.

국가 R&D 기반정보(이미경, 정한민, 성원경 2005; 이미경 외 2006)는 논문, 지적 재산권, 보고서를 포함하는 성과정보, 과제정보, 인력정보 등으로 구성되는, 국가 R&D 관련 응용서비스에서 반드시 참조(Reference)해야 하는 정보이다. OntoFrame K®와 같은 국가 R&D 기반정보에 기반한 응용 시스템을 구현하고자 할 때는 각 정보를 구성하는 인스턴스들에 있어서 애매성이 발생하면 안 된다(성원경, 정한민 2006). 예를 들어, 특정인력이 소속기관을 변경하거나 인적 사항을 변경하는 경우에도 일관된 접근방법을 제공하여야 하며, 동명이인 문제로 인해 문자열만으로 인력을 구분해서도 안 된다. 일관된 인스턴스의 식별을 위해서는 URI를 체계적으로 관리하고 서비스하는 시스템이 필요하다. 본 연구에서는 이러한 목적으로 URI 서버를 도입하고 있으며, 이를 통해 일관성 있는 정보등록 및 관리를 할 수 있도록 하며, 웹서비스를 이용하여 인터넷상에서 언제, 어디서나 정확한 URI를 제공할 수 있도록 한다.

의미표현이 결여된 현재의 웹은 지식단위의 인식과 해석의 어려움으로 인해 기계에 의한 자동처리가 쉽지 않다. 이러한 문제를 해결할 수 있는 차세대 웹의 비전으로 제시된 시맨틱 웹은, 웹 문서의 단위정보에 해당 분야 온톨로지의 개념 클래스로 정의된 의미 태그를 부착함으로써 이종 스키마를 갖는 정보들 간 의미적 통합 및 유통의 자동화와 명시적으로 표현되지 않은 암묵적 지식의 추론을 가능하게 한다. 특히, 시맨틱 웹의 성공을 위한 중요한 전제조건 중 하나는 개별 분야 온톨로지의 적절한 작성 및 왕성한 사용이다(Sure et al. 2005). 이러한 측면에서 본 연구는 국가 R&D 기반정보 온톨로지(이하 '기반정보 온톨로지'로도 명명함)를 설계하고, 그로부터 새롭게 얻어지는 지식의 추론과정을 통해 시맨틱 웹을 이용한 국가 R&D 기반정보를 체계적으로 서비스하고자 한다.

## 2. 관련 연구

온톨로지의 개별 개체(Individual)에 대해서는, 실세계에서의 그 인스턴스의 신원에 해당하는 고유한 URI를 대응시킬 필요가 있다. 이러한 인스턴스의 신원해소(Identity Resolution)를 위해서는, 인스턴스에 대한 언어적 표현의 동의성과 다의성 문제를 해소할 필요가 있다. 이 두 문제는 실세계의 인스턴스가 고유한 URI 대신 동의성과 다의성을 내재하고 있는 자연언어로 표현한다는 데서 기인한다. 예를 들면, 학술 데이터에서 인력 인스턴스

는 이름으로 표현되며, 기관 인스턴스는 고유한 기관 코드 값 대신 기관명으로 표현한다.

동일 인스턴스에 대한 서로 다른 표현문제인 동의성은 연구분야 온톨로지의 경우 그리 심각하지는 않다. 그 이유는 그들의 동의어 표현집합이 어느 정도 닫혀있다고 볼 수 있는 기관명이나 게재지명 등에 국한되기 때문이다. 그렇지만, 인력의 이름이 갖고 있는 서로 다른 인스턴스에 대한 형태적 동일 표현문제인 다의성은 실세계의 신원이 온톨로지 내에서 왜곡되는 결과를 초래하므로 반드시 해소되어야 한다. 본 연구에서는 이러한 동명이인 문제를 해소하기 위한 단서로서 동명 연구자의 문헌들에 내재된 공저자정보, 전자메일 주소, 소속기관명, 출판년도 등을 복합적으로 고려하는 방법을 개발하여 사용하고 있다. 이에 대해서는 논문에서 구체적으로 후술될 것이다.

한 문서 내에서의 인명에 대한 참조(Citation) 해결에 대해서는 이미 많은 연구가 있었지만, 여러 문서에 나타나는 동일 이름에 대한 참조해결에 대한 연구는 최근에 두드러지게 나타나고 있다. Mann 및 Yarowsky (2003)는 Bootstrapping을 통해 문서에서 자동으로 추출한 인물의 출생지, 생년월일, 직업 등의 사적인 기록들을 자질로 사용하는 비교사 (Unsupervised) 방식의 클러스터링 기법을 소개하였다. 이 방법은 동명이인을 두 클러스터로만 구분할 수 있다. Fleischman과 Hovy(2004)는 동명이인일 확률을 학습하는 Maximum Entropy 모델에 기반한 클러스터링 기법을 제시하였다. 최근에 Bekkerman 및 McCallum (2005)과 Malin(2005)은 사회망(Social Network)에서의 동명이인 구별문제를 다루었다. Bekkerman 및 McCallum(2005)은 서로 관련된 인물들의 이름과 매칭되는 웹 페이지에서 실제 그 인물들을 찾아내기 위해 링크 구조와 A/CDC 클러스터링 기법을 고안하였다. Malin (2005)은 인터넷 영화 DB상에서 동명이인을 구별하기 위해 이름에 기반한 관계 망의 유사성을 이용하였다.

이러한 연구들은 텍스트에서의 동명이인 문제를 다룬 것인데 반해, Alani(2002)는 시맨틱 웹을 위한 온톨로지의 통합과정에서 발생할 수 있는 동명이인 문제를 다루고 있다. 서로 다른 두 온톨로지에 나타나는 두 인물이 동일인인지 여부를 결정하기 위해 각 인물의 COP (Communities Of Practice)를 비교한다. COP의 유사성이 임계치 이상이면 동일인으로 간주한다. 본 연구에서 다루는 서지정보의 공저자 관계는 COP의 한 예라고 볼 수 있다는 점에서 유사성을 갖는다.

기존에 개발된 대표적인 연구분야 온톨로지로는 AKT 온톨로지, SWRC 온톨로지(Sure et al. 2005), Bibster 온톨로지(Haase et al. 2004)와 SWRC+COIN 온톨로지(Bloehdorn et al. 2005) 등이 있으며, 이들 연구분야 온톨로지 스키마를 구성하는 주요 클래스들은 인력, 기관, 과제, 논문 등이다. 기반정보 온톨로지는 온톨로지 스키마 측면에서는 이러한 기존

온톨로지들과 크게 다르지 않다. 그렇지만, 중요한 차이점 중 하나는 온톨로지에 표현될 인스턴스의 신원을 관리하기 위해 개별 클래스에 종속적인 URI 부여체계에 따라 고유한 인스턴스 ID를 할당한다는 데 있다. 인스턴스 ID의 생성, 저장, 검색 등의 관리를 위해 URI 서버가 사용된다는 것 또한 최초의 시도이다. URI 서버를 통해 인스턴스를 관리함으로써 데이터 정합성검사를 효율적으로 수행할 수 있다는 장점을 가진다.

기존의 연구분야 온톨로지들에서 찾아보기 힘든 본 기반정보 온톨로지의 또 다른 특징은 논문작성 당시의 저자 소속기관을 표현하기 위해 '저작자정보'라는 클래스를 도입한 것이다. 이렇게 함으로써 저자순위정보를 명시적으로 표현할 수 있고, 저자의 성과물작성 당시 기관과 현재 소속기관을 모두 표현함으로써 기관별, 개인별 논문실적을 정확히 산정할 수 있는 장점을 가진다.

## 3. 국가 R&D 기반정보의 구축

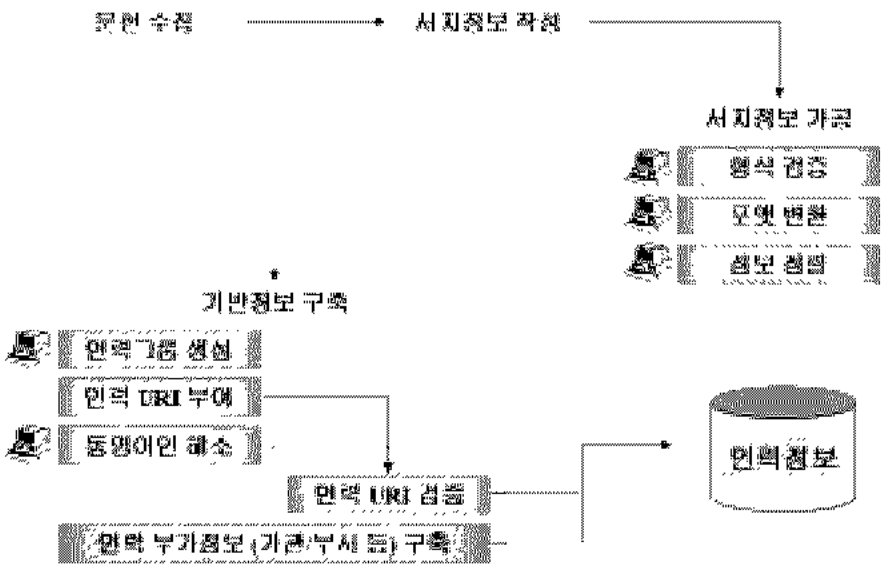
국가 R&D 기반정보는 인력정보, 과제정보, 성과정보, 그리고 기관, 부서 등을 포함하는 기타 정보로 구성되며, 이들은 다양한 식별체계들과 관계를 가진다(4.1절 참조). 본 연구는 이 중 인력정보, 성과정보, 기타 정보를 주로 다루고 있으며, 이들을 획득하기 위한 방안으로서 과학기술문헌을 사용한다[3]. 과학기술문헌 중 논문은 보고서, 지적 재산권과 더불어 국가 R&D 기반정보 중 성과정보를 구성하는 주요 요소이다. 본 장에서는 과학기술 인력정보를 포함하는 기반정보구축을 위한 기본자료로서의 서지정보를 작성하고 가공하는 방법과 URI를 부여하는 방법(<그림 1> 참조)을 설명한다.

### 3.1 과학기술문헌 서지정보의 작성 및 가공

#### 3.1.1 과학기술문헌 서지정보의 작성

논문으로 대표되는 과학기술문헌에 대한 서지정보는 제목, 저자정보(주저자, 소속기관, 소속부서, Email 등), 출처정보(문헌유형, 주최기관, 발행기관, 문헌명칭 등), 발행연도로 구성된다. 각 문헌마다 저자의 창작특성이 다르기 때문에 일관성 있는 서지정보 확보를 위한 입력지침이 필요하다. 예를 들어, 한글 저자명은 붙여 써야한다거나, 한글 소속기관명과 소속부서명은 따로 구분하되, 각각은 붙여 써야한다는 것 등이다. 특히, 학술대회 논문집에서는 소속정보를 기재하지 않거나, Email을 기재하지 않는 경우가 종종 발생한다. 소속정보나 Email을 기재하더라도 공저자를 포함한 전체 저자들과 매칭되지 않는 경우도 발생한다. 이러한 경

3) 과제정보는 KISTI 내부 성과정보와 연결된 유발과제들을 중심으로 구축하고 있으나, 본 논문의 기술범위에서는 제외한다.




〈그림 1〉 인력정보 구축 프로세스

우 위첨자를 분석하거나 Email의 이니셜로부터 유추하는 휴리스틱들을 사용하고 있으나 그 처리에 한계가 있을 수밖에 없다. 소속정보의 경우에도 "한국과학기술원", "과학기술원", "과기원", "KAIST"와 같이 다양한 표현들이 존재할 수 있으므로, 기관 URI를 이용하여 해결할 필요가 있다.

3.1.2 과학기술문헌 서지정보 가공

과학기술문헌 서지정보 입력 후 효율적인 URI 부여를 위해 서지정보를 가공할 필요가 있다. 즉, 같은 저자명을 빈도순으로 정렬함으로써 구축 우선순위 결정과 상호 비교를 통해 동명이인 문제해소가 용이해진다. 본 연구는 이러한 서지정보 가공을 형식검증, 포맷 변환, 정보정렬 과정으로 나눈다. 형식검증에서는 공백 문자, 줄 바꿈 기호 등의 불필요한 문자들을 제거하고 불완전한 저자정보를 재확인한다. 포맷 변환은 출처 URI, 인력 URI, 기관 URI, 부서 URI 등을 용이하게 구축할 수 있도록 하는 준비작업으로, 공저자 분석을 위해 공저자 목록 필드를 추가한다. 정보정렬은 한정된 인적·물적 자원을 이용하여 효율적으로 동명이인 문제를 해소하고자 구축목적에 맞도록 특정 필드 중심으로 정렬하는 작업이다. 본 연구에서는 동일한 저자명이 많은 순으로 정렬함으로써 인력정보 구축의 효율성을 극대화하고자 하며, 우선 구축대상으로 동명이인 문제가 발생하는 2번 이상 출현한 인명으로 한다. 다음은 가공된 서지정보가 가지는 필드들을 나열한 결

| | |
|---|---|
| **KISTI1.PCD.0003883** | **// ID (KOI 생성규칙 응용)** |
| 지능형로봇 환경에서의 질의처리 | // 논문 제목 |
| **2005HCI** | **// 출처 (HCI 2005년도 학술대회)** |
| docsrc\논문(원문및텍스트)\2005HCI\원문\00219 | // 파일명 |
| pdf | // 파일 타입 |
| 5 | // 총 저자 수 |
| 1 | // 저자 순위 |
| **정한민** | **// 저자명** |
| **한국과학기술정보연구원** | **// 소속기관** |
| **차세대정보시스템연구실** | **// 소속부서** |
| jhm@kisti.re.kr | // Email |
| 정한민;선충녕;손주찬;성원경;박동인; | // 공저자 목록 |

과와 그 예이다. 이 중 URI 부여대상은 볼드체로 표현한 ID, 출처, 저자명, 소속기관, 소속부서이다.

### 3.2 저자 그룹 생성 및 인력 URI 부여

인력 URI는 주민등록번호와 같이 유일한 식별자를 포함하여야 한다. 본 연구에서는 신규로 인력 URI를 설계하는 대신에 이미 실제 시스템에 적용된 식별자[4]를 이용하고자, 국가과학기술인력 종합정보시스템에서 사용하고 있는 인력 ID를 인력 URI로서 사용한다. 또한, 해당 URI를 발견할 수 없는 미등록 인력에 대해서는 동일한 인력 ID 생성규칙을 통해 가상 인력 ID를 생성하여 인력 URI로 사용한다.

#### 3.2.1 국가과학기술인력 종합정보시스템

국가과학기술인력 종합정보시스템은 국내 여러 기관에 분산되어 구축 운영 중인 인력 DB를 연계하여 통합 메타 DB를 구축하고, 통합검색 서비스 및 각종 현황정보 서비스를 제공하는 시스템이다. 현재 한국과학기술정보연구원(KISTI), 한국과학재단(KOSEF)을 포함하여 총 24개 기관이 참여하고 있다. 2006년 5월 현재 대학교, 연구소, 산업체 인력 등을 포함하여 약 32만7,000명이 등록되어 있다. 국가과학기술인력 종합정보시스템에서는 10자리로 구성된 고유 식별체계(첫 두 자리는 생년, 세 번째 자리는 성별, 나머지는 일련번호)를 인력 ID로 사용하고 있으며, 인명, 소장처 및 식별체계를 결합하여 해당 인력에 대한 상세정보에 바로

4) 본 논문에서는 URI를 구성하는 Namespace("http://www.kisti.re.isri#"), Prefix(각 URI 유형에 따른 3글자의 영문명과 ":"로 구성), 식별자(Identifier) 중 식별자만을 편의상 URI라고 표현하기도 한다. 그렇지만, 본 연구에서 설계 및 구현한 모든 식별자는 URI 기반으로 표현한다.

접근할 수 있도록 되어 있다. 예를 들어, 인명이 "정한민", 소장처가 "KISTI", 식별번호가 "7010186243"인 경우에 "http://hrst.or.kr/hrst/viewDetailFrameSet.jsp?koi=7010186243&korggubun=KISTI&kname=정한민"이라는 상세정보 URL을 자동생성할 수 있다. 본 연구에서는 이미 등록된 인력에 대해서는 국가과학기술인력 종합정보시스템의 인력 ID를 그대로 URI화하고, 미등록 인력에 대해서는 10자리를 유지하되 가상의 할당 영역에 해당할 수 있도록 "000"으로 시작하는 일련번호를 부여한다(예를 들어, 첫 번째 미등록 인력에는 "0000000001"을 식별번호로 부여하므로, 인력 URI는 "http://www.kisti.re.kr/isrl#PER_0000000001"이다).

3.2.2 인력 URI 획득 및 생성

인력 URI 획득 및 생성 프로세스는 다음 프로세스와 같다. 가공된 과학기술문헌 서지정보를 이용하여 저자 그룹들을 생성하고, 각 그룹에 대해 국가과학기술인력 종합정보시스템을 참조하여 인력 URI를 부여하거나 신규 인력 URI를 생성하여 할당한다.

다음은 이러한 프로세스에 의해 인력 URI를 해소한 결과 예를 보여준다. "조성배"는 세 인력 그룹으로 구분되며, 이중 두 그룹("6510145983", "7510237363")은 국가과학기술인력 종합정보시스템에 의해 인력 URI를 할당받고, 나머지 하나("0000000007")는 신규 인력 URI를 할당받는다.

1. 공저자 분석 및 그룹화: 동명이인 문제를 해소하고자 하는 인력에 대해 공저자를 공유하는 인력들을 동일 인물로 간주하여 그룹화 한다[5].
2. Email 분석 및 인력 그룹 병합: 두 인력 그룹이 같은 Email을 하나 이상 공유하고 있는 경우 두 그룹을 하나로 병합한다.
3. '소속+연도' 분석 및 인력 그룹 병합: 두 인력 그룹이 동일한 연도에 같은 소속(소속기관 및 소속부서)을 공유하고 있는 경우 두 그룹을 하나로 병합한다. 단, 소속정보가 불완전한 경우에는 다른 소속에 속하는 것으로 간주한다.
4. 인력 그룹 병합 및 인력 URI 획득: 국가과학기술인력 종합정보시스템을 검색하여 경력정보 확인 등을 통해 소속 변경 등으로 두 인력 그룹이 동일하다고 판단하는 경우에는 해당 인력 그룹들을 병합하고, 각 인력 그룹에 대해 등록 인력 ID를 인력 URI로서 할당한다.
5. 인력 URI 생성: 국가과학기술인력 종합정보시스템에 등록되지 않은 인력이라고 판단하는 경우에 신규 인력 URI를 생성하여 할당한다.

5) 이러한 작업에는 저자와 공저자 모두에서 동명이인이 동시에 발생하는 경우가 거의 없다는 가정에 기반한다. 이 가정은 현재 검증 중에 있다.
6) 2003 KISSS : 2003년 한국정보과학회 춘계학술대회, 2002KISSF : 2002년 한국정보과학회 추계학술대회

과학기술정보 납본제도 운영실태

| ID | 출처 | 저자명 | 인력 ID | 소속 | Email | 공저자 목록 |
|---|---|---|---|---|---|---|
| KISTI.PC D.0005989 | 2003KISSS[6] | 조성배 | 6510145983 | 연세대학교 컴퓨터과학과 | sbcho@cs.yonsei.ac.kr | 한상준;조성배; |
| KISTI.PC D.0005998 | 2003KISSS | 조성배 | 6510145983 | 연세대학교 컴퓨터과학과 | sbcho@csai.yonsei.ac.kr | 박찬호;조성배; |
| KISTI.PC D.0006014 | 2003KISSS | 조성배 | 7510237363 | 한국전기 연구원 | sbcho@keri.re.kr | 하현석;황민태; 조성배;이재조; |
| KISTI.PC D.0000349 | 2002KISSF | 조성배 | 6510145983 | 연세대학교 컴퓨터과학과 | sbcho@cs.yonsei.ac.kr | 한상준;조성배; |
| KISTI.PC D.0006161 | 2003KISSS | 조성배 | 0000000007 | 순천향대학교정보기술공학부 | hopi@dkpower.com | 김동균;전병찬; 조성배;이상정; |
| KISTI.PC D.0007149 | 2003KISSS | 조성배 | 0000000007 | 순천향대학교정보기술공학부 | hopi@dkpower.com | 송재훈;조성배; 이상정; |
| KISTI.PC D.0007032 | 2003KISSS | 조성배 | 6510145983 | 연세대학교 컴퓨터과학과 | sbcho@cs.yonsei.ac.kr | 민현정;김경중; 조성배; |

### 3.3 온톨로지 관계(Property) 이용

상기 작업절차에 의해 인력정보 구축작업을 수행할 때 완전하게 동명이인 문제가 해소되지 않는 경우가 발생할 수 있다. 예를 들어, 하나의 인력이 소속을 변경하여 논문을 작성한 경우에는 공저자, Email, 소속+연도 모두에 대해 불일치할 수 있다. 또한, 국가과학기술인력 종합정보시스템에도 해당 인력의 경력이 제대로 업데이트되지 않을 수 있다. 이러한 경우, 다른 인력으로 간주할 수밖에 없는데, 추후 경력정보가 업데이트가 되거나, 관련 홈페이지 등 다른 루트를 통해 동명이인 문제가 해소되는 상황을 가정해야 한다.

본 연구에서는 이러한 가능성을 현 시점에서 무리하게 해소하는 대신에 OWL(Web Ontology Language) 기반의 온톨로지 관계(Property)를 이용하여 해소되는 시점에 해소결과를 반영할 수 있도록 한다. OWL Lite와 OWL DL에서 제공하는 동등관계(Equality Property)들 중에서 'sameAs'는 서로 다른 이름을 가진 개체(Individual)들이 동일한 Resource를 지칭하고 있다는 의미를 부여한다. 이 관계(Property)를 이용하면 서로 다른 URI들을 가지는 개체(Individual)들을 하나로 연결할 수 있으며, 비록 두 인력이 서로 다른 URI들을 가지더라도 추론과정에서 동일인력으로 처리하는 것이 가능하다. 다음은 두 인력 ID를 'sameAs' 관계(Property)로 처리한 예이다.

```
〈Person rdf:ID="0000000169"〉
 〈owl:sameAs rdf:resource="#5910089763" /〉
  〈/Person〉
```

### 3.4 공저자 관계 분석 및 인력 URI 검증

논문 서지정보를 온톨로지로 표현하기 위해서는 서지정보에 포함된 각 개체를 유일하게 식별할 수 있어야 한다. 여기서 가장 문제가 되는 것이 저자명에 대한 동명이인 구별이다. 즉, 서로 다른 서지정보에 같은 이름의 저자가 나타날 때, 그것이 동일인을 가리키는 지 여부를 판단하는 것이 중요하다.

일반 문서에서와는 달리, 논문의 서지정보는 각 저자의 소속과 Email에 대한 정보를 포함하고 있어 비교적 동명이인 문제해소가 쉬울 수 있다. 그렇지만, 현실에서는 이 정보만으로 온전히 구분하기 어려운 경우가 흔히 발생한다.

첫 번째로 논문에 소속이나 Email 정보가 기재되지 않은 경우가 그러하다. 양식이 엄격한 논문지와는 달리, 학술대회 논문집인 경우, 비록 지정된 양식에는 소속과 Email을 기재하도록 되어 있다고 하더라도 저자가 이를 따르지 않는 경우가 흔히 발생한다. 저자가 다수인 경우, 각 저자별로 소속과 Email을 명확히 구분하지 않은 경우도 이에 해당한다.

두 번째로 동일 저자의 소속과 Email 표기가 일관되지 않은 경우이다. 소속을 표기하는 데 있어서, 기관만을 표기하거나 기관과 부서를 함께 표기한 경우, 부서를 'XYZ학과' 혹은 'XYZ학부', 'XYZ연구실' 과 같이 다르게 표현한 경우를 많이 발견할 수 있다. 규모가 큰 기관에서는 한 기관 내에 동명이인이 발생할 가능성이 크다는 점을 감안할 때, 저자의 기관명만으로 동명이인 문제를 해소하기에는 위험이 크다. 예를 들어, "연세대학교 컴퓨터과학과"의 "조성배"와 "연세대학교 대학원 인지과학협동과정"의 "조성배"가 동일인이라고 단정할 수 없다. Email의 경우도 마찬가지로, 한 사람이 두 개 이상의 Email 주소를 사용하는 경우도 흔히 볼 수 있다. 기관에서 제공하는 Email과 부서 혹은 연구실에서 제공하는 Email, 웹 포털 사이트에서 제공하는 Email 등 복수 개의 Email을 혼용하여 사용할 때, 그 Email이 가리키는 사람이 동일인임을 알기 어렵다. 예를 들어, "sbcho@cs.yonsei.ac.kr", "sbcho@sclab.yonsei.ac.kr", "sbcho@candy.yonsei.ac.kr", "sbcho@csai.yonsei.ac.kr" 과 같이 다른 Email을 갖는 여러 "조성배"가 동일인이라고 단정하기는 어렵다.

세 번째로 시간에 따라 소속과 Email이 바뀌는 경우이다. 저자가 소속을 옮기는 경우도 흔히 발생하며, 이때, Email도 함께 바뀌는 경우가 많다. 기관 내에서 부서를 옮기거나 부서명 자체가 변경되는 경우도 생각할 수 있다. 이 경우, 논문의 발행연도를 참조하는 것이 도움이 될 수 있으나 절대적이지는 않다.

이와 같이, 소속과 Email은 동명이인을 구별하는 데 필요한 중요한 정보를 제공하지만, 이것만으로는 부족한 경우가 흔히 발생하기 때문에 추가적으로 공저자 관계를 이용해야 한다.

#### 3.4.1 인력 URI 검증

인력 URI를 검증하기 위해 먼저 공저자 관

계 분석과 소속, Email을 이용하여 자동으로 인력 그룹을 생성한다.

Step 1. 소속(기관 및 부서), Email이 모두 일치하면 동일인으로 간주하여 하나의 인력 그룹 으로 묶는다. 하나도 없으면 멈춘다.

Step 2. 동일 이름의 공저자가 있으면 동일인으로 간주하여 해당 인력 그룹에 추가하며 소 속 집합, Email 집합을 구성한다. 하나 이상 있으면 Step 1로 간다. 하나도 없으면 멈 춘다.

상기 방법에 의해 작성된 인력 그룹 예는 〈표 1〉과 같다.

자동으로 생성된 인력 그룹과 구축자에 의해 수작업으로 만들어진 인력 그룹 간의 불일치를 발견하고, 이들에 대해 수작업으로 검증하여 오류를 수정하는 순으로 인력 URI 검증을 수행한다. 〈표 2〉는 두 인력 그룹 간의 불일치 결과 예를 보여준다. 'Disagreement#' 이 0이 아닌 경우가 불일치된 인력 그룹이 있음을

〈표 1〉 자동생성된 인력 그룹 예

| | | | | | |
|---|---|---|---|---|---|
| 이재호 | ID1 | 한국전자통신연구원 | 기반기술연구소 임베디드S/W기술센터 | bigleap@etri.re.kr | 이환구;김우식;이상윤;이재호;김선자; |
| 이재호 | ID1 | 한국전자통신연구원 | 한국전자통신연구원무선인터넷플랫폼팀 | bigleap@etri.re.kr | 김유일;이원재;한환수;이재호;김선자; |
| 이재호 | ID1 | 한국전자통신연구원 | 인터넷정보가전연구부 | bigleap@etri.re.kr | 이재호;김선자; |
| 이재호 | ID1 | 한국전자통신연구원 | 인터넷정보가전연구부 | bigleap@etri.re.kr | 이재호;김홍남;이현철; |
| 이재호 | ID1 | 한국전자통신연구원 | 인터넷정보가전연구부 | bigleap@etri.re.kr | 이재호;김선자;김성조;선동국;심형용; |
| 이재호 | ID10 | | | | 이재호;남광우;김민수; |
| 이재호 | ID10 | | | | 이재호;남광우;심우성; |
| 이재호 | ID11 | | | | 이재호;임덕성;홍봉희; |
| 이재호 | ID12 | 한국전자통신연구원 | 열차제어연구그룹 | | 김용규;백종현;이재호;유창근; |
| 이재호 | ID13 | 한국항공대학교 | 항공전자 및 정보통신공학 레이더 신호처리 연구실 | | 곽영길;전인평;최민수;황광연;이강훈;이재호; |
| 이재호 | ID2 | 서울시립대학교 | 전자전기컴퓨터공학부 | jaeho@ee.uos.ac.kr | 이경록;김인철;이재호; |
| 이재호 | ID3 | 한국방송공사 | 기술연구소 | jaeho@kbs.co.kr | 김진우;이재호;김희정; |
| 이재호 | ID4 | 서울시립대학교 | 전자전기컴퓨터공학부 | jaeho@uos.ac.kr | 곽별샘;변무홍;이재호; |
| 이재호 | ID5 | | | jaeho1@ysc.ac.kr | 이재호;이윤수;윤경섭;왕창종; |

. . . . . .

〈표 2〉 자동생성 인력 그룹과 수작업 인력 그룹 간의 불일치 예

| Name | Total# | Disagreement# | Over-clustering# | Under-clustering# |
|---|---|---|---|---|
| 이철훈 | 990 | 0 | 0 | 0 |
| 김상욱 | 741 | 0 | 0 | 0 |
| 김성수 | 630 | 0 | 0 | 0 |
| 신동일 | 595 | 0 | 0 | 0 |
| 류근호 | 351 | 27 | 0 | 27 |
| 유혁 | 303 | 48 | 0 | 48 |
| 이만재 | 254 | 46 | 0 | 46 |
| 김일곤 | 300 | 0 | 0 | 0 |
| 이재호 | 300 | 0 | 0 | 0 |
| 정순기 | 300 | 0 | 0 | 0 |
| 백두권 | 235 | 41 | 0 | 41 |
| 이상훈 | 274 | 2 | 1 | 1 |
| 최명렬 | 253 | 0 | 0 | 0 |
| 박형우 | 68 | 163 | 0 | 163 |
| ..... | | | | |

〈표 3〉 인력 URI 검증 예

| | | | | |
|---|---|---|---|---|
| 김민수 | 0000000095 | 0000000093 | | |
| 이병호 | 0000000322 | 0000000321 | 가톨릭대학교컴퓨터 | 정보공학부 |
| 이종원 | 0000000428 | 0000000426 | 서울대학교 | 컴퓨터공학부 |
| 이성호 | 0000000910 | 00000009 | 13Electronics and Telecommunications Research Institute | |
| 이정수 | 0000000920 | 0000000917 | | |
| 이승룡 | 0000000421 | 5510060767 | | |
| 이승룡 | 0000000421 | 5510060767 | 경희대학교 | 컴퓨터공학과 |
| 권오경 | 0000000438 | 5510061663 | 한국과학기술정보연구원 | 슈퍼컴퓨팅센터 |
| 권오경 | 0000000439 | 5510061663 | 한양대학교 | 전자통신컴퓨터공학부 |
| 이건표 | 0000000203 | 5610066835 | 한국과학기술원 | 산업디자인학과 |
| 박종원 | 0000000473 | 5610069878 | 충남대학교 | 정보통신공학과 |
| 차의영 | 0000000215 | 5610072400 | 부산대학교 | 멀티미디어협동과정 |
| 홍봉희 | 0000000169 | 5910089763 | | |
| ..... | | | | |

의미하며, 해당 그룹 내의 모든 인력 쌍에 대해서 비교했을 때의 불일치 수가 'Disagreement#'에 해당한다.

불일치 인력 그룹에 대해 수작업으로 검증한 결과 전체 1만8,200 명 중 72 명이 수정되어 약 0.4%의 수작업 오류를 발견할 수 있었다(<표 3>). 이 결과는 3.3의 온톨로지 관계(Property)를 이용하여 처리한다.

### 3.5 인력정보 및 성과정보 구축결과

인력정보와 성과정보를 포함하는 기반정보 구축을 위해 수집·구축한 과학기술문헌 종류 및 크기는 <표 4>와 같다. 구축 크기가 수집 크기와 다른 것은 중복 논문들과 원본을 얻을 수 없는 논문들을 배제했기 때문이다.

중복을 포함하여 전체 저자 출현 횟수는 2만 3,105건(논문 당 평균 공저자수는 약 3.22명)이며, 한 논문의 최대 공저자 수는 17명이다. 동명이인을 포함하여 동일명 저자의 최대 출현 횟수는 55회이며, 4,903명은 1회만 출현한 저자들이다. 또한, 소속이 기재되지 않은 저자는 27회, Email이 기재되지 않은 저자는 5,562회, 소속과 Email 모두 기재되지 않은 경우도 16회나 출현하였다. 본 구축에서는 동명이인 문제해소를 위해 2번 이상 성과물에서 나타난

<표 4> 과학기술문헌 구축결과

| 학회명 | 연도 | 학술대회명 | 크기 (구축/수집) |
|---|---|---|---|
| 한국정보과학회 | 2002 | 춘계학술대회 | 555/591 |
| | | 추계학술대회 | 757/7682003 |
| | 2003 | 춘계학술대회 | 774/794 |
| | | 추계학술대회 | 870/87 |
| 대한전자공학회 | 2004 | 춘계학술대회 | 682/729 |
| | 2003 | 하계학술대회 | 665/721 |
| | 2004 | 하계학술대회 | 419/421 |
| | 2005 | 추계학술대회 | 290/316 |
| 한국 HCI 학회 | 2003 | 학술대회 | 241/253 |
| | 2004 | 학술대회 | 312/318 |
| | 2005 | 학술대회 | 326/346 |
| | 2006 | 학술대회 | 369/383 |
| 한국정보처리학회 | 2004 | 추계학술대회 | 484/485 |
| | 2005 | 춘계학술대회 | 431/432 |
| 총합 | | | 7175/7431 |

〈표 5〉 인력 URI 획득 및 생성 결과

| 내 용 | 건수 |
|---|---|
| 물 수(- 동명이인 수) | 18,199 |
| Distinct 성과물 수 | 6,825 |
| 동일 인명 수 | 4,369 |
| Distinct 인력 수 | 6,976 |
| 국가과학기술인력 종합정보시스템 등록 인력 수(Distinct 인력 수 – 신규 URI 할당 인력 수) | 1,176 |

인력들을 대상으로 하였다. 〈그림 1〉에서 정의한 인력 URI 획득 및 생성 작업절차를 통해 구축한 결과는 〈표 5〉와 같다.

### 3.6 기타정보 구축결과

기관정보를 위해서 한국학술진흥재단을 중심으로 KISTI에서 수집 · 통합한 1만2,111개의 기관 코드를 식별자로서 사용하였다. 기관 코드 및 기관명 외에 각 기관정보를 웹 검색하여 우편번호로 구성된 위치정보와 홈페이지 정보를 추가하였다. 이외에 성과정보에서 추가로 발견된 218개 기관들을 포함시켰다.

부서정보를 위해 한국학술진흥재단을 중심으로 KISTI에서 수집 · 통합한 4,745개 부서 코드를 식별자로서 사용하였다. 부서의 경우 성과정보에서 정확히 명칭이 매칭되지 않는 경우들이 많아 부서 코드 할당에 다음과 같은 원칙을 세워 작업을 수행하였다.

정확히 일치하는 부서명을 사용한다.

"학과 · 학부"까지 일치하지 않는 경우 "학"까지 일치하는 부서명을 사용한다(예. "컴퓨터공학").

"과 · 부"까지 일치하지 않는 경우 "과 · 부"가 생략된 부서명을 사용한다(예. "일본어")

상기 해당 사항이 없는 경우 신규 등록한다.

## 4. URI 관리 및 서비스

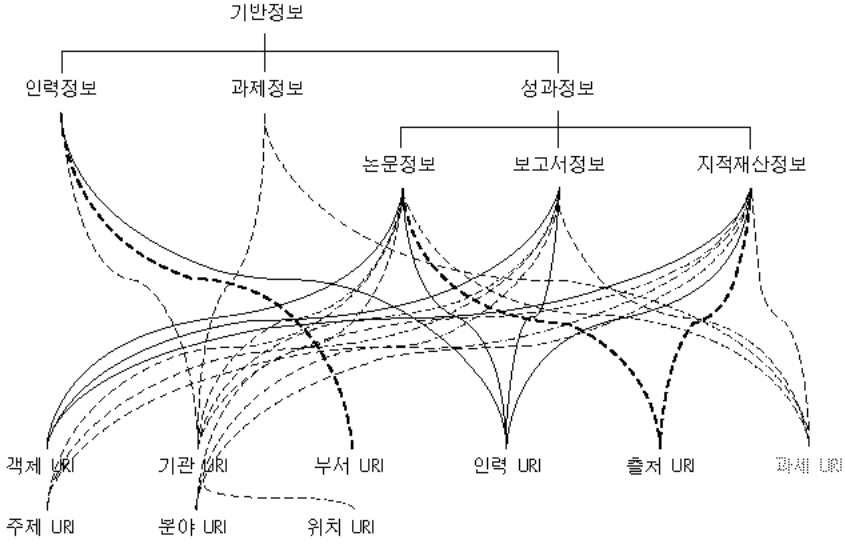
국가 R&D 기반정보 온톨로지는 크게 인력정보, 과제정보, 성과정보로 구성된다. 이들은 여러 URI들을 포함하는 메타데이터로 표현할 수 있다. 4장에서는 이러한 특성을 반영한 다중 URI 관리 및 서비스를 포함하는 URI 서버를 설명한다. URI 서버는 관리 관점의 웹 기반 URI 관리 시스템과 서비스 관점의 웹 서비스 기반 URI 서비스 시스템으로 구성된다. 또한, 기존정보는 이미 구축 · 활용되고 있는 식별체계(한국과학기술정보연구원의 KOI, 한국학술진흥재단의 기관 및 부서 코드, 국가과학기술인력 종합정보시스템의 인력 ID 등)를 도입하고 신규정보에 대해서는 URI 생성규칙을 제공한다.

### 4.1 국가 R&D 기반정보 온톨로지상의 URI

〈그림 2〉는 국가 R&D 기반정보 온톨로지를 구성하고 있는 인력정보, 과제정보, 성과정보와 각종 URI들(객체 URI, 기관 URI, 부서 URI, 인력 URI, 출처 URI, 과제 URI, 주제 URI, 분야 URI 및 위치 URI) 간의 관계를 보여준다. 성과정보 중 하나인 보고서를 예로 들어 설명하면, 보고서는 그 보고서를 유발케 한 과제 URI, 보고서 자체를 표현하는 객체 URI, 보고서 작성자들의 인력 URI, 해당 인력의 소속기관을 의미하는 기관 URI, 그리고 보고서의 주제 및 분야에 해당하는 주제 URI와 분야 URI 등을 포함한다.

URI는 해당 Identity를 유일하게 지칭할 수 있는 식별자의 일종으로 네임스페이스(Namespace), Prefix, 식별자(identifier)로 구성된다. 인력 URI를 예로 들어 설명을 하면, "http://www.kisti.re.kr/isrl#PER_7010186243"과 같다. 네임스페이스는 이름을 구분짓기 위한 추상적 그룹의 집합체이고, Prefix는 식별자부여 정책충돌을 해소하기 위한 구분자이다. 일반적으로 Prefix를 사용하지 않는데, 본 연구에서는 기존 식별체계들을 도입하다 보니 이들 간의 식별자가 충돌하는 경우가 발생하므로(예. 기관과 부서) 이의 해결을 위해 Prefix를 사용한다. 다음은 URI 구성요소 중 핵심이 되는 식별자에 대한 생성규칙을 설명한다.

객체 식별자는 호환성을 고려해 한국과학기술정보연구원의 KOI 생성규칙에 기반하여 정의한다. 생성규칙은 "등록관리기관코드.자료유형코드.일련번호"로 구성된다. '등록관리기관코드'는 객체의 종류에 따라 학술지, 보고서,




〈그림 2〉 온톨로지의 구성요소 및 URI 연관도

개인공유자료, 기반정보 등으로 구분하며, '자료유형 코드'는 학술지, 학술대회, 학위논문, 특허, 연구보고서, 개인공유자료 등으로 구분한다. 마지막으로 '일련번호'는 7자리 숫자를 사용한다.

기관 식별자는 한국학술진흥재단의 4~6자리 영어 대문자와 숫자가 조합된 식별체계를 사용하며, 신규기관에 대해서는 첫 자리를 0으로 갖는 6자리 숫자를 사용한다.

부서 식별자 역시 한국학술진흥재단의 4자리 숫자로 구성된 식별체계를 사용하며, 신규부서에 대해서는 4자리 숫자로 표현하는데, 첫 1자리를 "S"로 표현하고, 나머지 3자리는 숫자를 사용한다.

인력 식별자는 국가과학기술인력 종합정보시스템의 식별체계를 따르는데, 국가과학기술인력 종합정보시스템에 등록된 인력은 10자리 인력 ID를 사용하고, 등록되지 않은 인력 역시 10자리를 사용하되, 첫 3자리를 "0"으로 설정하여 기존인력 ID와의 충돌을 방지한다.

출처 식별자는 출처식별코드 3자리와 일련번호 6자리로 구성되며, 출처식별코드는 논문지의 경우 "SOJ", 학술대회의 경우 "SOP", 학위논문의 경우 "SOT", 지적 재산권의 경우 "SOI", 보고서의 경우 과제 URI를 사용한다.

과제 식별자는 영숫자 혼용 형태로 한국과

〈표 6〉 식별자 예제

| | 식별자 예제 | |
|---|---|---|
| 객체 식별자 | KISTI.SOJ.000001<br>KISTI.PTN.000001 | //지식기반구축 플랫폼 연구<br>//지식기반구축 방법론 특허 |
| 기관 식별자 | 8A5327<br>9R9048<br>000002 | //한국정보과학회<br>//한국과학기술정보연구원<br>//오롬정보 |
| 부서 식별자 | 5459<br>S001 | //전기, 전자, 컴퓨터공학<br>//NTIS사업단 |
| 인력 식별자 | 6410136403<br>7010186243<br>0000003451 | //성원경<br>//정한민<br>//구희관 |
| 출처 식별자 | SOP000001<br>SOJ000001<br>SOT000003 | //2005년정보과학회 추계학술대회<br>//정보과학회 논문지, 시스템 및 이론<br>//과학기술연합대학원대학교 박사학위 |
| 과제 식별자 | Z-2006-008919<br>Z-2006-008830 | //해외과학기술정보 수집분석 활용사업<br>//슈퍼컴퓨팅 인프라 기술 연구 |
| 주제 식별자 | GE0305 | // 범용 과학기술 분야 "선형 가속기" |
| 분야 식별자 | 020302 | // 그래픽스 |
| 위치 식별자 | 305-806 | // 대전광역시 유성구 어은동 1~99 |

학기술정보연구원의 과제번호 부여 체계를 따른다. 외부 과제에 대해서는 "Z 연도 일련번호"로 구성하며, 연도는 4자리, 일련번호는 6자리 숫자를 사용한다.

주제 식별자는 과학기술 분야식별코드 2자리와 일련번호 7자리로 구성된다. 과학기술 분야식별코드는 범용과학기술의 경우 "GE", IT의 경우 "IT"로 정의한다.

분야 식별자는 짝수 개의 숫자로 구성된다. 각 2자리 숫자는 분야분류체계에서의 1 Depth이다. 예를 들어, "030213"라는 식별자는 "03"이 대분류, "02"가 중분류, "13"이 소분류임을 의미한다.

위치 식별자를 위해서는 우편번호를 그대로 사용한다.

### 4.2 URI 관리 및 서비스 시스템

#### 4.2.1 URI 관리 및 서비스 시스템 설계원칙

URI 관리 및 서비스 시스템은 서비스 레이어(Service Layer), 비즈니스 레이어(Business Layer), 퍼시스턴스 레이어(Persistence Layer)의 3계층으로 구성된다. 이중 서비스 레이어는 크게 관리 측면과 서비스 측면으로 분리하여 설계한다. 관리 측면의 설계는 기존 URI 데이터의 입력 및 확인 작업을 위한 웹 인터페이스로 구현되며 이를 URI 관리 시스템이라 하고, 서비스 측면의 설계는 다양한 클라이언트에게 서비스를 제공할 수 있는 웹 서비스 인터페이스로 구현되며 이를 URI 서비스 시스템이라 한다.

#### 4.2.2 URI 관리 및 서비스 시스템의 비즈니스 레이어

상기 두 가지 측면의 인터페이스는 공통의 컴포넌트인 비즈니스 레이어를 이용하여 URI 관리 및 서비스를 제공한다. 비즈니스 레이어는 다중 URI에 대해 일관된 접근을 제공하며 두 개의 인터페이스에서 발생한 같은 자원에 대한 요구를 효과적으로 중재한다.

비즈니스 레이어가 제공하는 URI 등록은 URI 요청, 확정, 취소 및 형식 유효성 검사 등 4가지 기본적인 메소드(Method)로 구성된다. URI 확정은 기존정보를 1차적으로 검색하여 그 일치 여부에 따라 등록을 결정하는 것을 의미하며, URI 오류 및 정보 중복검사 등을 위한 검증을 포함한다. 정보 중복검사는 정보별 검사항목 일치 여부를 판별하는 방식으로 이루어진다. 예를 들어, 인력정보 중복검사는 인력 · 기관 · 부서 URI 및 Email 일치여부를 검사한다.

URI 등록을 작업 순서로 구분하면 등록요청, 서버 응답, 사용자확인의 3단계로 나눌 수 있다. 등록의 효율성을 증가시키기 위해 타임스탬프를 URI 등록요청 시에 비즈니스 레이어에서 생성한다. 한 번 생성이 되고 나서 사용자 확인 없이 정해진 시간을 넘기는 경우에는 생성대기 큐에 반납된다.

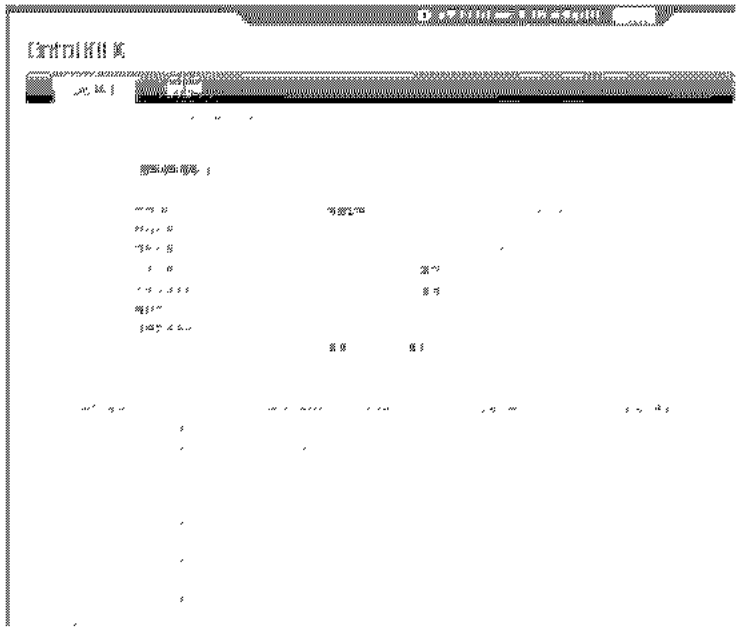
결과적으로 비즈니스 레이어를 공통으로 사용하게 함으로써 모든 자원관리가 일관성을 지

니게 된다. URI의 확정과 취소과정이 특정 인터페이스에서 발생하더라도 유효한 URI를 분배하고 반영할 수 있다.

4.2.3 URI 관리 시스템

사용자를 위한 관리 측면의 웹 인터페이스 기반 URI 관리 시스템은 서비스 레이어에서 동작한다. 이미 할당된 URI들의 반입 및 관리를 담당하며 사용자권한에 따라 기능을 세분화한다.

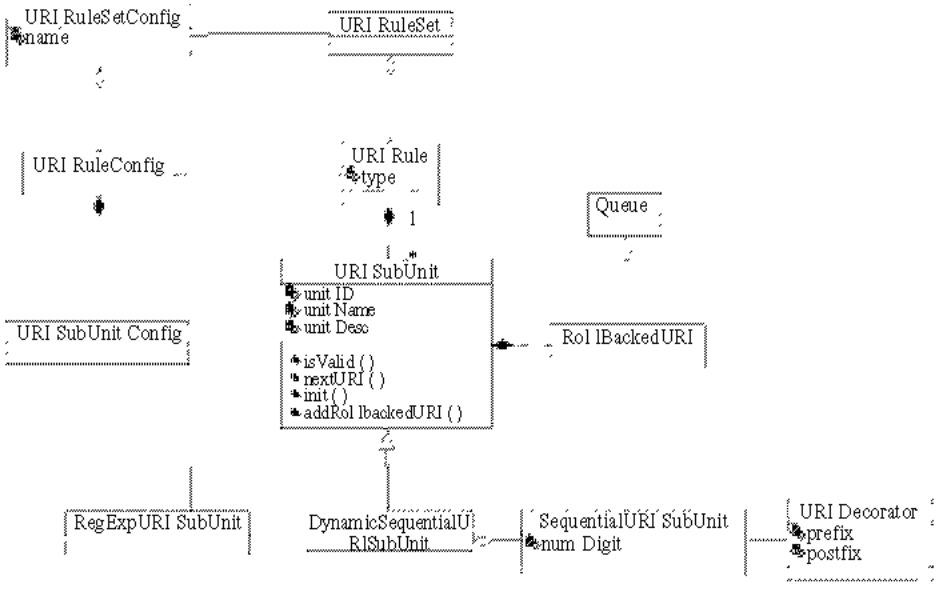
〈그림 3〉은 3장에서 얻어진 인력정보들을 관리화면을 이용하여 등록한 결과를 보여준다. 인력정보등록 화면은 인력정보 URI를 직접 입력할 수 있는 인력 URI 필드와 한글이름 필드, 영문이름 필드, 소속기관 URI 필드, 소속부서 URI 필드, 메일주소 필드, 상세정보 URL 필드



〈그림 3〉 인력 URI 관리화면(조회 및 등록)

로 구성된다. 등록화면 하단에는 등록된 인력정보를 보여준다(실제 구현에서는 '인력 URI 조회'에서 볼 수 있다.).

또한, 인명 등록 시에 소속기관 URI를 필수적으로 참조하도록 함으로써 데이터의 신뢰성을 증가시킨다. URI 생성규칙에 의거해 인력 URI를 자동생성하거나 직접 입력하는 경우에도 URI 생성규칙을 이용해 검증한다.




〈그림 4〉 URI 관리 시스템 클래스 구조도

4.2.4 URI 서비스 시스템

서비스 측면의 웹 서비스 인터페이스 기반 URI 서비스 시스템 역시 서비스 레이어에서 동작하며 다중 URI를 동시에 구별하여 처리할 수 있도록 분할 설계한다. 그 이유는 웹 서비스 특성상 응답속도의 저하가 발생할 수 있기 때문에 URI 서비스 시스템의 연산을 최소화하기 위해 URI 요청에서 클라이언트가 URI를 유형

〈표 7〉 URI 웹 서비스 이름 및 대상 URI

| 서비스 이름 | 대상 URI | 명세 |
|---|---|---|
| ArtCom_Service | 객체 URI | requestURI, commitURI, rollbackURI, validateURI의 4개의 등록관련 연산과 일반조회 및 상세 결과조회에 관한 연산으로 구성 |
| OrgURI_Service | 기관 URI | |
| DepURI_Service | 부서 URI | |
| HumURI_Service | 인력 URI | |
| ConCom_Service | 출처 URI | |
| ProURI_Service | 과제 URI | |
| TopURI_Service | 주제 URI | |
| CatURI_Service | 분야 URI | |
| LocURI_Service | 위치 URI | |

별로 구분하여 요청하도록 설계하는 것이다 (<표 7> 참조).

URI 서비스 시스템은 URI 관리 시스템을 통해 구축된 자원을 다양한 클라이언트들에게 제공하기 위한 목적으로 설계되었으며, 공통의 비즈니스 레이어가 제공하는 메소드들을 모두 사용할 수 있다.

## 5. 국가 R&D 기반정보 온톨로지 설계

### 5.1 온톨로지 개요 및 설계 원칙

기반정보 온톨로지는 과학기술분야의 연구·개발자들의 협업 연구를 가상공간에서 지원하기 위한 소프트 인트라의 하나로 개발되고 있다. 현재의 기반정보 온톨로지는 논문, 과제, 지적 재산권 등의 연구·개발 성과물을 중심으로, 관련된 인력, 기관, 연구분야·주제 등을 모델링하고 있다.

기반정보 온톨로지 설계의 중요한 원칙으로는 시나리오 지향성, 간결성, 그리고 URI 서버 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(Derivable) 객체 관계(Object Property)들을 온톨로지 스키마에 포함시키지 않는다는 것을 의미한다. URI 서버 지향성은 클래스 인스턴스의 고유성 보장을 위해 개별 클래스에 종속적인 URI 생성규칙에 따라 인스턴스 ID를 부여한다는 것과 클래스에 의존적인 속성 관계(Datatype Property)들은 온톨로지 스키마에 표현하지 않고 URI 서버에서 관리하도록 한다는 것을 의미한다. URI 서버를 통한 관리는 전술한 온톨로지의 간결성 원칙과도 부합되는 것이다.

온톨로지 기술 언어로는 W3C의 시맨틱 웹 표준 언어인 OWL을 사용하였고, 온톨로지 편집도구로는 Protege 3.1.1을 이용하였다.

### 5.2 온톨로지 스키마

과학기술 연구·개발 분야를 다루는 기반정보 온톨로지는 현재 핵심적인 최상위 클래스로 인력, 기관, 과제, 저작물, 게재지, 주제를 포함하고 있다. 저작물은 논문, 특허, 연구보고서의 하위 클래스들로 세분되며, 게재지는 학술지(논문지)와 학술대회 논문집의 하위 클래스를 가진다. 상기 핵심 상위 클래스들 간의 객체관계 속성을 최소한으로 나열하면 다음과 같다.

저작물 hasCreatorsInformation(저작자 정보를 갖는다) 저작자정보

저작물 hasOriginatedProject(유발과제를 갖는다) 과제

저작물 hasPublication(게재지를 갖는다) 게재지

저작물 hasTopic(주제를 갖는다) 주제

저작자정보 hasCreator(저작자를 갖는다) 인력

저작자정보 hasOrganizationOfCreator(저작 당시 기관을 가진다) 기관

과제 hasOrganizationOfFundingProject(발주기관을 가진다) 기관

과제 hasOrganizationOfPerformingProject(수탁기관을 가진다) 기관

주제 hasSubTopic(하위주제를 가진다) 주제

인력 hasOrganizationOfPerson(현재 소속기관을 가진다) 기관

저작자정보라는 핵심 상위 클래스가 있는데, 이것은 저작물의 특정 순위 저자를 모델링한 것으로 저작물 작성 당시의 저자가 갖는 속성들을 포함한다. 저작자정보는 데이터 타입 속성으로 저자순위, 저자기여도를 가지며, 저작물, 인력, 기관 등의 클래스와 객체관계 속성을 맺고 있다. 즉, 한 저작물의 특정 순위 저자에 해당하는 인력을 표현하기 위해 저작자정보와 인력 간에 객체관계를 설정하고, 특정 순위 저자의 저작물 작성 당시의 소속기관을 표현하기 위해 저작자정보와 기관 간에 객체관계를 설정한 것이다.

기반정보 온톨로지 스키마에서 특별히 주목할 점은 그것의 개별 클래스들이 데이터 타입 속성을 거의 갖지 않는다는 것이다. 예를 들어, 기반정보 온톨로지에서 인력 클래스는 인력의 이름과 같은 기본적인 데이터 타입 속성을 갖지 않으며, 과제 클래스의 경우도 과제의 명칭과 같은 데이터 타입 속성이 표현되어 있지 않다. 이러한 데이터 타입 속성들은 특정 클래스에 종속적인 것들인데, 기반정보 온톨로지에서는 이러한 데이터 타입 속성들을 URI 서버에서 관리하는 구조를 갖도록 최초로 설계되었다. URI 서버와 온톨로지 사이는 URI를 매개로 연결된다(5.3절의 예제 참조).

기반정보 온톨로지는 연구개발 분야에서 기존에 개발된 대표적인 온톨로지들인 KA2 온톨로지, AKT 온톨로지, SWRC 온톨로지들과 비교했을 때, 온톨로지 모델링 관점에서는 유사한 점을 가진다. 그렇지만, 기반정보 온톨로지는 인력이나 기관의 연구성과를 정확히 산정하기 위해 기존 온톨로지에서는 찾아볼 수 없는 저작자정보라는 클래스를 도입한 점과 인스턴스의 고유한 표현을 위해 동명이인 문제를 해소하는 전처리 과정을 거친다는 특징을 가진다.

### 5.3 기반정보 온톨로지 인스턴스 표현 예

〈그림 5〉는 기반정보 온톨로지 스키마에 대응하는 인스턴스 표현의 한 예를 보여준다.

[과제]

과제명: 생물 유전자원정보 네트워크 구축 및 관련기술연구

과제기간: 2004~2005

발주·수탁기관: 과학기술부, 한국과학기술정보연구원

[과제의 유발성과물]

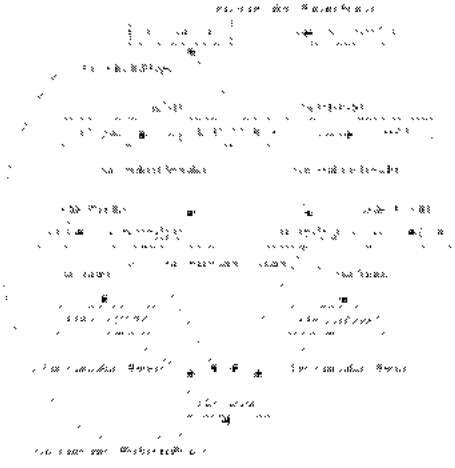
논문명: 생물다양성 데이터 활용을 위한 국가 데이터 노드 구축

저자: 박형선, 안성수, ...

게재지명: 지식정보인프라

주제: 유전자, ...

〈그림 5〉에서 개별 인스턴스들은 그것의 고유한 URI로 표현되고 있다. 기반정보 온톨로지에서 각 클래스들은 인스턴스 ID를 할당하는 URI 생성규칙과 관련된다. 예를 들어, 저작물 인스턴스는 KOI 식별체계를 간략화한 규칙에 따라 URI가 부여되고, 인력 인스턴스에는 국가과학기술인력 종합정보시스템에서 사용하는 인력식별체계를 근간으로 하는 URI가 할당된다. 〈그림 5〉에서 알 수 있듯이 클래스 종속적



| URI | 유형 | 메타데이터 |
|---|---|---|
| PER:6110113732 | 인력 | 이름:박형선, 소속기관: ORG:9R9048 |
| PER:0000000001 | 인력 | 이름:안성수, 소속기관: ORG:9R9048 |
| OBJ:KISTM.JNL.0000003 | 저작물 (논문) | 제목:생물 다양성 데이터 활용을 위한 국가 데미터노드 구축<br>관련과제:PRO:G-04-TJ-103 |
| PRO:G-04-TJ-103 | 과제 | 제목:생물, 유전자원정보 네트워크 구축 및 관련기술 연구<br>연도:2004-2005, 발주기관:ORG:9T9187,<br>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 5〉 URI 지향적 기반정보 온톨로지 인스턴스 표현 예
(표 내용은 URI 서버에 저장된 데이터이며, URI에서 편의상 식별자만 보여준다.)

```
  <CreationInformation rdf:ID="저작자정보_012">
  <orderOfCreator rdf:datatype="http://www.w3.org/2001/XMLSchema#int"
  >1</orderOfCreator>
  <hasOrganizationOfCreator>
    <Organization rdf:ID="ORG_9R9048" />
  </hasOrganizationOfCreator>
  <hasCreator>
    <Person rdf:ID="PER_6110113732">
      <hasOrganizationOfPerson rdf:resource="#ORG_9R9048" />
    </Person>
  </hasCreator>
</CreationInformation>
<Journal rdf:ID="PUB_SOJ000913" />
<Person rdf:ID="PER_0000000001">
  <hasOrganizationOfPerson rdf:resource="#ORG_9R9048" />
</Person>
<Organization rdf:ID="ORG_9T9187" />
<CreationInformation rdf:ID="저작자정보_013">
  <hasOrganizationOfCreator rdf:resource="#ORG_9R9048" />
  <hasCreator rdf:resource="#PER_0000000001" />
  <orderOfCreator rdf:datatype="http://www.w3.org/2001/XMLSchema#int"
  >2</orderOfCreator>
</CreationInformation>
<Project rdf:ID="PRO_G-04-TJ-I03">
  <hasOrganizationOfPerformingProject rdf:resource="#ORG_9R9048" />
  <hasOrganizationOfFundingProject rdf:resource="#ORG_9T9187" />
</Project>
<Paper rdf:ID="OBJ_KISTI1.JNL.0000003">
  <hasCreationInformation rdf:resource="#저작자정보_013" />
  <hasCreationInformation rdf:resource="#저작자정보_012" />
  <hasOriginatedProject rdf:resource="#PRO_G-04-TJ-I03" />
  <hasTopic>
    <Topic rdf:ID="CAT_120403" />
  </hasTopic>
</Paper>
```

인 데이터 타입 속성 값들은 거의 온톨로지에 표현되지 않으며, URI 서버에서 관리된다. 온톨로지와 URI 서버와의 연결은 개별 인스턴스의 URI로 가능하다. 〈그림 5〉를 OWL (Ontology Web Language)로 기술한 예가 이어진다.

## 6. 추론 서비스

시맨틱 웹상에서 온톨로지는 다양한 이질적 데이터 스키마를 가진 데이터 소스들을 의미적으로 통합하고 유통시킬 수 있을 뿐만 아니라, 선언적 규칙을 사용하여 온톨로지 내에 명시적으로 표현되지 않은 암묵적 연관관계들을 추론하는 용도로 사용될 수 있다. 이 장에서는 연구분야 온톨로지의 한 예로 제시된 5장의 기반정보 온톨로지로부터 규칙을 적용하여 온톨로지에 내재된 지식들을 추론하는 과정을 기술한다. 이를 위해 먼저 기반정보 온톨로지에서 예상되는 규칙들을 정의하고 그 규칙에서 기술된 객체관계속성을 이용하여 기반정보 온톨로지에 내재된 객체 간의 관계들을 추출하는 RDQL(RDF Data Query Language) 질의의 예를 보인다.

다음은 기반정보 온톨로지에 내재된 지식들을 추출하는 몇 가지 규칙들을 Jena 추론 엔진의 규칙기술형식에 따라 쓴 것이다. 각 규칙은 → 를 중심으로 하여 IF THEN 형식으로 표현한 것이며, Rule 2에 사용된 notEqual( )은 일반적인 추론 엔진들이 지원하는 built in 함수로서 그것의 두 매개변수가 같지 않다는 것을 의미한다. 각 규칙과 의미는 다음과 같다.

Rule 1: (?x hasCreatorsInformation ?y) (?y hasCreator ?z) → (?x wasCreatedBy ?z)

Rule 2: (?x wasCreatedBy ?y) (?x wasCreatedBy ?z) notEqual(?y, ?z) → (?y isCoCreatorOf ?z)

Rule 3: (?x hasCreatorsInformation ?y) (?y hasCreator ?z) (?y orderOfCreator 1) → (?z isFirstCreatorOf ?x)

Rule 4: (?x hasCreatorsInformation ?y) (?y hasOrganizationOfCreator ?z) → (?x isOwnedByOrganization ?z)

Rule 5 1: (?x isCoCreatorOf ?y) → (?x isSameGroupWith ?y)

Rule 5 2: (?x isSameGroupWith ?y) (?y isSameGroupWith ?z) → (?x isSameGroupWith ?z)

Rule 1: ?x라는 저작물이 ?z라는 인력을 저자로 갖는 어떤 저작자정보 ?y를 가진다면, 그 저작물 ?x는 인력 ?z에 의해 작성된 것이다

Rule 2: 서로 다른 두 인력 ?y와 ?z가 동일한 저작물 ?x를 작성했다면 ?y와 ?z는 각각 서로의 공저자이다

Rule 3: ?x라는 저작물이 ?z라는 인력을 제1저자로 갖는 어떤 저작자정보 ?y를 가진다면, 저작물 ?x의 제1저자는 ?z이다.

Rule 4: ?x라는 저작물이 어떤 인력이 ?z라는

기관에 소속되었을 당시에 작성한 저작물이 라면, 저작물 ?x는 기관 ?z의 실적이다.

Rule 5 1: 공저자 관계에 있는 두 인력은 같은 연구자 그룹에 속한다.

Rule 5 2: ?y와 같은 연구자 그룹에 속하는 서로 다른 두 인력 ?x와 ?z는 역시 같은 연구자 그룹에 속한다.

규칙 5 1과 5 2에서는 저작물에 대해 단순히 직간접적 공저자 관계에 있는 연구자들을 동일 연구자 그룹의 구성원으로 고려하고 있다. 위의 규칙들을 기반정보 온톨로지에 적용함으로써 아래와 같은 새로운 객체관계속성들을 얻을 수 있다.

저작물 wasCreatedBy 인력

인력 isCoCreatorOf 인력

인력 isFirstCreatorOf 저작물

저작물 isOwnedByOrganization 기관

인력 isSameGroupWith 인력

다음은, 위의 규칙들이 적용되어 확장된 기반정보 온톨로지에서 데이터베이스 분야의 연구자들을 인력 URI 형식으로 추출하는 RDQL 질의 예를 보여준다.

```
SELECT      ?y
WHERE       (?x wasCreatedBy ?y)
(?x hasTopic "데이터베이스")
```

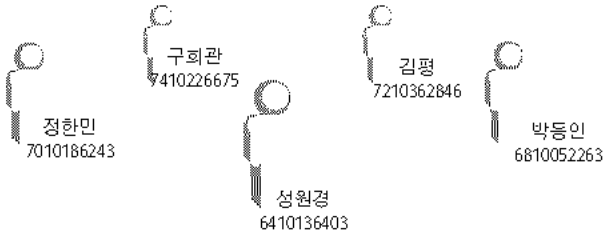
다음 RDQL질의는, 확장된 온톨로지에서 URI "4410022529"를 갖는 인력("이석호", 〈그림 4〉 참조)과 같은 연구자 그룹에 속하는 인력들을 인력 URI 형식으로 출력해 준다.

```
SELECT      ?y
WHERE       ("4410022529"
isSameGroupWith ?y)
```

〈그림 6〉은 기반정보 온톨로지로부터 추론될 수 있는 연구자 네트워크에 대한 사용자 인터페이스의 한 예를 보여주고 있다. 인력과 인력 사이의 연결선은 두 인력이 공동으로 연구한 성과물(논문, 특허, 과제 등)의 유무를 표시하며, 연결선의 굵기는 협업 성과물의 양적·질적 강도[7]를 나타낸다. 전면 중앙의 인력을 중심으로 하는 연구자 네트워크에서 마우스 클릭을 통해 다른 인력을 중심으로 하는 새로운 연구자 네트워크를 동적으로 생성할 수 있다. 그림에 표시되지 않은, 각 인력과 관련된 인력 간의 숨겨진 연결선들을 MORE 버튼을 클릭함으로써 검색할 수 있다.

기반정보 온톨로지의 스키마 부분은 5.2절에 기술된 핵심 상위 클래스를 포함하여 현재 20개의 클래스와 39개 속성들(데이터 타입 속성 20개, 객체관계 속성 19개)로 구성되어 있으며, 9개의 규칙이 기술되어 있다. 현재 기반정보 온톨로지의 인스턴스 부분은 IT 분야 6,825건의 논문들로부터 구축되어 있다. 위 논문들은 〈표 4〉에 보인 것처럼 2002년부터 2006년까지 한국정보과학회, 대한전자공학회,

7) 논문의 경우 논문지, 학술대회 논문집 등 유형에 따라, 그리고 국내, 국외와 같은 출처 등에 따라 각기 다른 가중치를 가지고 있으므로 이를 활용하여 질적 강도를 구할 수 있다.




〈그림 6〉 기반정보 온톨로지로부터 추론되는 연구자 네트워크의 예

〈표 8〉 기반정보 온톨로지 구축현황

| 구 분 | | RDF Triple 개수 |
|---|---|---|
| 클래스 인스턴스 | 인 력 | 6,976 |
| | 저작물 (논문) | 6,825 |
| | 기 관 | 308 |
| | 부 서 | 548 |
| | 게재지 (학술대회논문집) | 14 |
| | 저작자정보 | 18,200 |
| | 주 제 | 942 |
| | 합 계 | 33,813 |
| 인스턴스 간 객체관계 | | 128,342 |
| 총 합 | | 162,155 |

HCI학회, 한국정보처리학회 등에서 주최한 학술대회 논문집에서 추출한 것들이다. 구축된 인스턴스는, 규칙의 적용을 통해 얻어지는 Triple을 제외하고, 총 16만2,155개의 RDF Triple들로 구성되어 있다. 표에서 알 수 있듯이 현재 이 Triple 개수에는 과제 · 주제 클래스와 관련된 인스턴스들은 포함되어 있지 않다.

## 7. 결 론

본 연구는 실제 과학기술문헌들을 대상으로 하여 URI 기반으로 인력 및 성과 중심의 국가 R&D 기반정보를 구축하고, 이를 이용하여 추론을 통한 응용 서비스를 제공하는 방안을 설명하였다. 이러한 URI 기반 표현은 의미 기반 정보검색, 추론 서비스와 같이 문자열만으로는 해결할 수 없는 고급 응용 서비스 분야에서 필수적으로 요구되는 것으로, 온톨로지상의 인력, 객체, 기관, 부서, 출처 등 다양한 개체

(Individual) 유형들에 동시에 적용한 최초의 시도였다는 데 그 의의가 있다. 본 연구를 통해 구현하는 국가 R&D 기반정보 응용 서비스는 연구자 간 협업 및 과학기술 정책결정을 위한 기본정보제공 등의 다양한 영역에서 그 가치를 인정받을 것이다.

## 참고문헌


성원경, 정한민. 2006. OntoFrame K®: 협업 연구 지원을 위한 시맨틱 웹 기반 지식 정보 공유·유통 플랫폼. 『정보과학회지』, 24(4).

신동구, 김재수, 윤정모, 권이남, 전성진, 정택영. 2005. 식별체계 간 연계 시스템 구축에 관한 연구. 『한국정보과학회 추계학술대회』.

이미경, 정한민, 성원경. 2005. 지식 기반정보 유통 플랫폼 개발. 『제10회 한국과학기술정보인프라 워크숍』.

이미경, 정한민, 이승우, 성원경. 2006. 국가과학기술 R&D 기반정보 온톨로지 구축 및 적용. 『한국정보처리학회 춘계학술대회』.

이상환, 신동구, 김재수, 최진영, 정택영. 2004. 식별체계 기반의 과학기술분야 전자원문 연계시스템 설계 및 구현. 『한국정보과학회 춘계학술대회』.

장대근, 지수영, 오원근, 김의정. 1998. 논문 첫 페이지 영상의 논리적 구조형성을 위한 제목, 저자, 요약 영역의 추출. 『한국정보처리학회 추계학술대회』.

Alani, H., Dasmahapatra, S., Gibbins, N., Glaser, H., Harris, S., Kalfoglou, Y., O'Hara K., and Shadbolt, N. 2002. "Managing Reference: Ensuring Referential Integrity of Ontologies for the Semantic Web." *Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management.*

Bekkerman R. and McCallum, A. 2005. "Disambiguating Web Appearances of People in a Social Network." *Proceedings of the WWW Conference.*

Bloehdorn, S., Haase, P., Hefke, M., Sure, Y., and Tempich, C. 2005. "Intelligent Community Lifecycle Support." *Proceedings of the 5th International Conference on Knowledge Management.*

Fleischman M. and Hovy, E. 2004. "Multi Document Person Name Resolution." *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.*

Haase, P., Broekstra, J., Ehrig, M.,

Menken, M., Mika, P., Olko, M., Plechawski, M., Pyszlak, P., Schnizler, B., Siebes, R., Staab, S., and Tempich, C. 2004. "Bibster A Semantics Based Bibliographic Peer to Peer System." *Proceedings of the 3rd International Semantic Web Conference*.

Harth, A. 2004. "SECO: Mediation Services for Semantic Web Data", *Journal of IEEE Intelligent Systems*. 19(3).

Malin, B. 2005. "Unsupervised Name Disambiguation via Social Network Similarity." *Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security, in Conjunction with the SIAM International Conference on Data Mining*.

Mann, G. and Yarowsky D. 2003. "Unsupervised Personal Name Disambiguation." *Proceedings of the 7th Conference on Computational Natural Language Learning*.

Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., and Oberle, D. 2005. "The SWRC Ontology Semantic Web for Research Communities." *Proceedings of the 12th Portuguese Conference on Artificial Intelligence*.