과학기술 문헌으로부터의 URI 기반 인력정보 구축

The Construction of URI-Based Human Resource Information from Science and Technology Papers

정한민, 이승우, 강인수, 성원경
한국과학기술정보연구원 정보시스템연구팀

Han-Min Jung(jhm@kisti.re.kr), Seung-Woo Lee(swlee@kisti.re.kr),
In-Su Kang(dbaisk@kisti.re.kr), Won-Kyung Sung(wksung@kisti.re.kr)

요약

시맨틱 웹의 발전은 온톨로지를 포함한 언어 자원들에 기초하고 있으며, 온톨로지 상에서 개체(Individual)들은 식별체계인 URI(Universal Resource Identifier)를 이용하여 유일하게 지칭될 수 있도록 구축되어야 한다. 그렇지만, 현실에서 식별체계를 사용하는 경우를 발견하기가 힘들며, 특히 논문과 같은 과학기술 문헌은 그 적용 대상에서 제외되어 왔다. 이러한 이유로 인해 과학기술 문헌상의 인력정보를 식별체계 기반으로 구축하고자 하는 시도가 미약한 실정이었다. 이에 본 논문은 과학기술 문헌으로부터 인력정보를 내부와 외부로 나누어 URI 기반으로 구축하는 방법을 기술한다. 이 때, 인력정보 자동 검증 방법을 적용하여 구축 초기에 참고정보를 제공하거나 구축 후에 인력정보를 검증할 수 있도록 한다. 본 논문은 공저자 관계, 전자우편, 발행년도, 소속기관 등을 이용하여 동명이인 문제를 해소하고, 각 저자 그룹 별 URI 부여를 위해 국가과학기술인력 종합정보시스템을 활용한 사례를 소개하는 방식으로 기술한다. 이러한 과정을 통해 9,484건의 과학기술 문헌들로부터 획득한 외부 인력정보와 KISTI 내부 인력정보는 연구자 네트워크 분석, 성과 통계 등 다양한 시맨틱 웹 응용 분야들에 필수적으로 활용될 것이다.

■ 중심어 : | URI(Uniform Resource Identifier) | 인력정보 | 온톨로지 | 시맨틱 웹 |

Abstract

The development of Semantic Web basically requires knowledge induced from the formalization and semantization of information, and thus ontology should be introduced as a knowledgization tool. URI(Universal Resource Identifier) is an indispensible scheme to uniquely indicate individuals on ontology. However, it is difficult to find the use cases of URI in real data including science and technology papers. This paper describes the method to construct internal and external human resource information based on URI from the papers. We use co-authors, e-mails, publication date, and affiliation for discriminating authors with the same strings. HRST(Human Resources devoted to Science and Technology) is referred to acquire URIs for human resource. We expect the internal and external human resource information would be adopted to outcome analysis applications such as researcher network analysis and outcome statistics.

■ keyword : | URI(Uniform Resource Identifier) | Human Resources | Ontology | Semantic Web |

접수번호 : #060719-002
접수일자 : 2006년 07월 19일
심사완료일 : 2006년 08월 25일
교신저자 : 강인수, e-mail : dbaisk@kisti.re.kr

I. 서 론

시맨틱 웹의 발전과 함께 온톨로지 구축의 중요성이 점점 강조되고 있는 현실에서, 온톨로지 표현의 기반이 되는 URI(Universal Resource Identifier)가 개체(Individual)들을 위한 식별체계로서 필수적으로 요구된다. 그렇지만, 이러한 중요성에도 불구하고 연구실 수준의 실험 데이터가 아닌 실제 데이터로부터 정보를 추출하고 정보 간 충돌을 해소한 후 URI를 부여하고자 하는 시도가 외국의 몇몇 사례들을 제외하고는 찾아보기 힘든 형편이다[9]. [8]에서와 같이 인력정보에 대해 식별자 기반으로 서비스를 제공하는 곳이 있기는 하나 해당정보는 실 데이터로부터 획득된 것이 아닌 사용자가 직접 등록한 정보에 불과하고 식별체계 적용에 있어서도 한계를 가지고 있다. 실 데이터로부터 발생할 수 있는 다양한 문제들을 해결할 수 있는 방안을 제시하지 않고서는 지속적인 정보 확장이 어려울 수밖에 없는 이유가 여기에 있다.

본 논문에서 추구하는 인력정보 구축은 결국 국가 과학기술 기반정보 온톨로지의 핵심이 되는 정보로서 그 구축의 정확성이 보장되어야 하며, 지속적인 인력정보 추가 시에도 체계적인 관리와 정합성을 갖추어야 한다. 이러한 측면에서 구축의 상당 부분을 수작업에 의존할 수밖에 없으며, 여러 자동화된 기법을 보조적으로 사용하여 그 효율성을 높일 필요가 있다. 기존에 논문으로부터 저자정보를 포함한 서지정보를 자동 추출하여 구축하는 연구가 있었다[4]. 그렇지만, 저자정보의 의미적 구축이 아닌 문자 인식 기반 문자열 추출에 초점이 맞추어 졌을 뿐이다. 다만, 식별체계 구축 시스템 간의 연계에 대한 연구가 [2]와 [3]을 통해 조금씩 이루어지고 있는데, 본 연구에서 과학기술 문헌 개체에 부여하는 ID를 KOI(Knowledge Object Identifier) 기반으로 구축함으로써 이러한 연구 결과를 부분적으로 활용하고 있다[5][6].[1]

본 논문에서는 인력정보 구축의 외부(II장 참조)와 내부(III장 참조)로 나누어 설명한다. IV장에서는 서지정보와 텍스트 처리 기술을 이용하여 인력정보를 검증하는 방안을 소개하며, V장에서는 구축 과정에서는 발생할 수 있는 동일 인력에 대한 서로 다른 URI 부여를 어떻게 해결할 것인지를 설명한다. 마지막으로 VI장에서는 인력정보 구축 결과를, VII장에서는 본 연구의 결론을 기술한다.

1 성과정보에서 URI 부여 대상은 논문명, 소속기관, 소속부서, 저자 등 다양하지만, 본 논문에서는 인력 URI(저자)만을 다루므로, 나머지는 참고문헌을 참조하기 바란다.

II. 외부 인력정보 구축

과학기술 문헌은 보고서, 지적재산권과 더불어 국가 과학기술 R&D 기반정보 중 성과정보를 구성하는 주요 요소이다. 본 장에서는 과학기술 인력정보 중 외부 인력정보를 구축하는 방법을 설명한다.

1. 과학기술 문헌 서지정보 작성

과학기술 문헌 서지정보는 제목, 저자정보(주저자, 소속기관, 소속부서, 전자우편 등), 출처정보(문헌유형, 주최기관, 발행기관, 문헌명칭 등), 발행연도로 구성된다. 각 문헌마다 저자의 창작 특성이 다르기 때문에 일관성 있는 서지정보 확보를 위한 입력 지침이 필요하다. 예를 들어, 한글 저자명은 붙여 써야 한다거나 한글 소속기관명과 소속부서명은 따로 구분하되 각각은 붙여 써야 한다는 것 등이 있다. 특히, 학술대회 논문집에서는 소속정보를 기재하지 않거나, 전자우편을 기재하지 않는 경우가 종종 발생한다. 소속정보나 전자우편을 기재하더라도 공저자를 포함한 전체 저자들과 매칭되지 않는 경우도 발생한다. 이러한 경우 위첨자를 이용하거나 전자우편의 ID를 분석하는 휴리스틱들을 사용하고 있으나 그 처리에 한계가 있을 수밖에 없다. 소속정보의 경우에도 "한국과학기술원", "과학기술원", "KAIST"와 같이 다양한 형태들이 존재할 수 있는데, 기관 URI를 이용하여 해결할 필요가 있다.

2. 과학기술 문헌 서지정보 가공

과학기술 문헌 서지정보 입력 후 효율적인 인력정보 구축을 위해 서지정보를 가공할 필요가 있다. 즉, 같은 저자명을 출현빈도순으로 정렬함으로써 구축 우선순위

결정과 상호 비교를 통한 용이한 동명이인 문제 해소가 가능해진다. 본 논문은 이러한 서지정보 가공을 형식 검증, 포맷 변환, 정보 정렬 과정으로 나눈다. 형식 검증에서는 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 정지 작업으로, 공저자 관계 분석을 위해 공저자 목록 필드를 추가한다. 정보 정렬은 한정된 인적 물적 자원을 활용하여 효율적으로 동명이인 문제를 해소하고자 구축 목적에 맞도록 특정 필드 중심으로 정렬하는 작업이다. 본 논문에서는 동일한 저자명이 많은 순으로 정렬함으로써 인력정보 구축의 효율성을 극대화하고자 한다. 다음은 가공된 서지정보가 가지는 필드들을 나열한 결과와 그 예이다. 이 중 URI 부여 대상은 ID, 출처, 저자(주저자, 공저자), 소속기관, 소속부서이다.

```
KISTI1.PCD.0003883 // ID(KOI 생성규칙 응용)
지능형로봇 환경에서의 질의처리 // 논문명
2005HCI // 출처(HCI 2005년도 학술대회)
2005HCI\원문\00219 // 파일명
pdf // 파일 유형
5 // 총 저자 수
1 // 저자 순위
정한민 // 주저자
한국과학기술정보연구원 // 소속기관
차세대정보시스템연구실 // 소속부서
jhm@kisti.re.kr // 전자우편
정한민;선충녕;손주찬;성원경;박동인; // 공저자
```

3. 국가과학기술인력 종합정보시스템

국가과학기술인력 종합정보시스템[8]은 국내 여러 기관에 분산되어 구축 운영 중인 인력 DB를 연계하여 통합 메타 DB를 구축하고, 통합 검색 서비스 및 각종 현황정보 서비스를 제공하는 시스템이다. 현재 한국과학기술정보연구원(KISTI), 한국과학재단(KOSEF)을 포함하여 총 24개 기관이 참여하고 있다. 2006년 7월 현재 대학교, 연구소, 산업체 인력 등을 포함하여 약 327,000명이 등록되어 있다. 국가과학기술인력 종합정보시스템에서는 10자리로 구성된 고유 식별체계를 인력 ID로 사용하고 있으며, 인명, 소장처 및 식별체계를 결합하여 해당 인력에 대한 상세정보에 바로 접근할 수 있도록 해준다. 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 자동 생성할 수 있는 특징을 가진다. 본 연구에서는 이미 등록된 인력에 대해서는 국가과학기술인력 종합정보시스템의 식별번호를 그대로 URI화하고, 미등록 인력에 대해서는 10자리를 유지하되 가상의 할당 영역이 될 수 있도록 "000"으로 시작하는 일련번호를 부여한다(예를 들어, 첫 번째 미등록 인력에는 "0000000001"을 식별번호로 부여하므로, URI는 "http://www.kisti.re.kr/isrl#PER_0000000001"이 된다.).[2]

4. 인력 URI 획득 및 생성

외부 인력 URI 획득 및 생성 방법은 다음과 같이 정의된다. 가공된 과학기술 문헌 서지정보를 이용하여 저자 그룹들을 생성하고 각 그룹에 대해 국가과학기술인력 종합정보시스템을 참조하여 등록 인력 URI를 부여하거나 신규 인력 URI를 생성하여 부여한다.

1. 공저자 분석 및 그룹화: 동명이인 문제를 해소하고자 하는 인력에 대해 공저자를 공유하는 인력들을 동일 인물로 간주하여 그룹화한다.
2. 전자우편 분석 및 인력 그룹 병합: 두 인력 그룹이 같은 전자우편을 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 AND 소속부서)을 공유하고 있는 경우 두 그룹을 하나로 병합한다. 단, 소속정보가 불완전한 경우에는 같은 소속이 아

2 URI는 네임스페이스("http://www.kisti.re.kr/isrl#"), 식별유형("PER_"), 식별번호("0000000001")로 구성되는데, 식별유형은 온톨로지에 따라 생략 가능하다. 본 논문에서는 편의상 식별번호를 URI로 지칭하기도 한다.

닌 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 부여한다.
5. 인력 URI 생성: 국가과학기술인력 종합정보시스템에 등록되지 않은 인력이라고 판단하는 경우에 신규 인력 URI를 생성하여 부여한다.

다음은 상기 절차에 의해 동명이인 문제를 해소하고 인력 URI를 부여한 결과의 예를 보여준다. "조성배"는 세 인력 그룹으로 구분되며, 두 그룹("6510145983", "7510237363")은 국가과학기술인력 종합정보시스템에 의해 인력 URI를 부여받고, 나머지 하나("0000000007")는 신규 인력 URI를 부여받는다.

KISTI1.PCD.00059892003 KISSS 조성배 6510145983 연세대학교 컴퓨터과학과 sbcho@cs.yonsei.ac.kr 한상준;조성배;3
KISTI1.PCD.00059982003 KISSS 조성배 6510145983 연세대학교 컴퓨터과학과 sbcho@csai.yonsei.ac.kr 박찬호;조성배;
KISTI1.PCD.00060142003 KISSS 조성배 7510237363 한국전기연구원 sbcho@keri.re.kr 하현석;황민태;조성배;이재조;
KISTI1.PCD.00003492002 KISSF 조성배 6510145983 연세대학교 컴퓨터과학과 sbcho@cs.yonsei.ac.kr 한상준;조성배;
KISTI1.PCD.00061612003 KISSS 조성배 0000000007 순천향대학교 정보기술공학부 hopi@dkpower.com 김동균;전병찬;조성배;이상정;
KISTI1.PCD.00071492003 KISSS 조성배 0000000007 순천향대학교 정보기술공학부 hopi@dkpower.com 송재훈;조성배;이상정;
KISTI1.PCD.00070322003 KISSS 조성배 6510145983 연세대학교 컴퓨터과학과 sbcho@cs.yonsei.ac.kr 민현정;김경중;조성배;

III. 내부 인력정보 구축

내부 인력정보 구축을 위해 활용하는 내부 성과정보는 한국과학기술정보연구원(Korea Institute of Science and Technology: KISTI)의 소속 인력들이 연구 성과로 창출한 학술정보로서 KISTI 그룹웨어에 의해 관리된다. 내부와 외부를 구분하는 이유는 내부 기반정보의 경우 그것을 구성하는 핵심 요소인 과학기술 문헌, 연구보고서, 지적재산권을 포함하는 성과정보와 인력, 과제 등의 연계정보에 해당하는 원문과 메타데이터를 DBMS를 통해 쉽게 획득할 수 있기 때문이다. 즉, 내부 인력정보 구축 과정은 그룹웨어 또는 기타 콘텐츠 관리 시스템에 존재하는 메타데이터를 본 연구에서 정의한 서지정보 형식으로 매핑하고 정제하는 일련의 과정으로 볼 수 있으므로, 메타데이터부터 구축하는 외부 인력정보 구축 과정과는 차별화된다.

1. 내부 인력정보 구축 방안

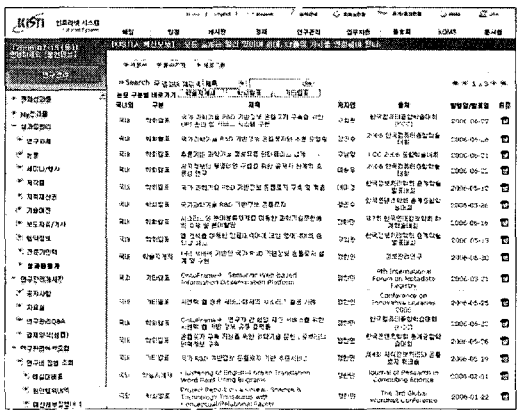

그림 1. KISTI 그룹웨어 내의 내부 성과정보 예

내부 인력에서도, 외부 인력의 경우보다 그 크기와 정도가 심하지는 않지만, 동명이인 문제가 발생한다. 그러나 일반적으로 기관이나 회사는 그 조직에 소속된 개별 인력에게 사번 또는 직번 등 자체적인 식별자를 부여하여 해당 인력들을 관리한다. 본 연구 대상인 KISTI 내부 성과정보에도 각 KISTI 소속 인력에 대응하는 직번이 포함되어 있다. 이러한 직번을 활용하면 동명이인 문제는 손쉽게 해결될 수 있다. 즉, 직번이 KISTI 내부 성과정보 내에서 고유하므로, 개별 직번마다 그에 대응되는 인력 URI로서 국가과학기술인력 종합정보시스템의 인력 URI를 사용하거나 인력 URI 생성 규칙에 맞게 신규로 부여하면 된다. 결국 KISTI 내부인력 중 국가과학기술인력 종합정보시스템에 등록된

3 "KISTI1.PCD.00059892003"은 KOI에 해당하는 URI이며, 소속기관과 소속부서 역시 URI 부여 대상이나 편의상 예제 기술에서 생략한다.

인력을 선별하는 작업을 선행해야 하며, KISTI 내부인력 중 국가과학기술인력 종합정보시스템에 미등록된 인력에 대해서만 신규 인력 URI를 부여하면 된다.

2. 내부 인력정보 구축

KISTI 내부 성과정보로부터 추출하여 구성하는 내부 인력정보는 다음의 6가지 항목으로 정의된다.

직번
인명
소속기관
소속부서
전자우편
인력 URI

위에서 인력 URI를 제외한 나머지 항목들은 KISTI 그룹웨어 내의 인력정보 관련 DB 테이블들을 대상으로 SQL(Structured Query Language) 문을 정의하고 실행시켜 추출할 수 있다. 물론 SQL 문의 작성을 위해서는 KISTI DB 테이블들의 스키마 구조를 이해하는 과정이 먼저 요구된다. [표 1]은 SQL 문을 통해 자동 추출된 초기 내부 인력정보의 예를 보여준다.

표 1. KISTI DB에서 추출된 내부 인력정보 초기 상태

직번	인명	소속기관	소속부서	전자우편	인력 URI
120040703	정한민	KISTI	정보시스템연구팀	jhm@kisti.re.kr	
120040101	성원경	KISTI	정보시스템연구팀	wksung@kisti.re.kr	
920060303	강인수	KISTI	정보시스템연구팀	dbaisk@kisti.re.kr	
920060301	이승우	KISTI	정보시스템연구팀	swlee@kisti.re.kr	

모든 내부인력의 소속기관은 "KISTI"로 동일하다.[4] 또한, 동일 인명이 존재하더라도 해당 조직 내에서 고유함이 보장되는 직번을 통해 그들의 신원을 쉽게 파악할 수 있다. 이를 통해 동일 인명으로 창출된 과학기술 문헌, 연구보고서, 지적재산권 등의 연구 성과물들을 직번 기준으로 구분한다.

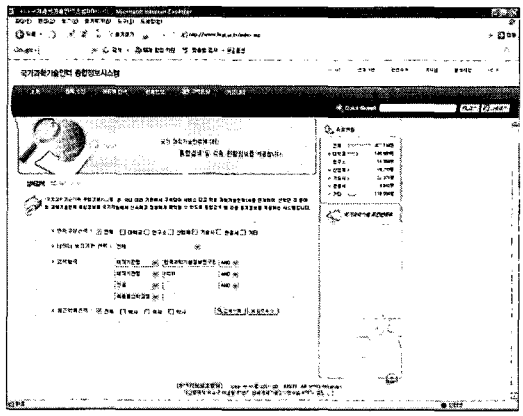

그림 2. 과학기술인력 종합정보시스템에서의 KISTI 내부 인력 검색 예

3. 인력 URI 부여

인력 URI 부여는 각각의 인력에 대해 과학기술인력 종합정보시스템에의 등록 여부를 확인하여, 등록된 경우 과학기술인력 종합정보시스템의 ID를 인력 URI로 부여하는 과정과 그렇지 않은 경우 신규 인력 URI를 부여 과정을 포함한다. KISTI 내부인력들이 과학기술인력 종합정보시스템에 등록되었는지를 확인하기 위해, 먼저 과학기술인력 종합정보시스템 홈페이지[8]에서 현재 재직기관이 "한국과학기술정보연구원"이거나 "KISTI"인 인력을 검색하고, 검색된 인력들을 표1에서 찾아 해당 인력의 인력 URI 항목을 채우는 방식으로 진행한다[그림 2][표 2].

내부 인력정보 구축의 마지막 단계는 과학기술인력 종합정보시스템에서 확인되지 않은 내부인력들에 신규 인력 URI를 부여하는 작업이다. 신규 인력 URI는 외부 인력정보에서 마지막으로 부여된 신규 인력 URI 이후 값을 순차적으로 사용한다[표 3][1].[5]

4 내부 성과정보에서도 외부 인력이 참여 연구원이나 공저자로서 나타날 수 있다. 본 연구에서는 내부 성과정보에 출현한 외부 인력은 내부 인력정보 구축 대상에서 제외하였다.

5 본 연구에서는 URI 관리를 위해 URI 서버를 별도로 구성한다. URI 서버는 웹 서비스 기반으로 시스템이 요청한 식별유형에 맞는 URI를 생성하여 서비스한다.

표 2. 과학기술인력 종합정보시스템으로부터의 등록된 인력들에 대해 인력 URI를 부여한 후의 내부 인력정보 상태(회색 배경은 과학기술인력 종합정보시스템에 등록된 인력)

직번	인명	소속기관	소속부서	전자우편	인력 URI
120040703	정한민	KISTI	정보시스템연구팀	jhm@kisti.re.kr	7010186243
120040101	성원경	KISTI	정보시스템연구팀	wksung@kisti.re.kr	6410136403
920060303	강인수	KISTI	정보시스템연구팀	dbaisk@kisti.re.kr	
920060301	이승우	KISTI	정보시스템연구팀	swlee@kisti.re.kr	

표 3. 신규 인력 URI 부여 후의 내부 인력정보 상태(회색 배경은 신규 인력)

직번	인명	소속기관	소속부서	전자우편	인력 URI
120040703	정한민	KISTI	정보시스템연구팀	jhm@kisti.re.kr	7010186243
120040101	성원경	KISTI	정보시스템연구팀	wksung@kisti.re.kr	6410136403
920060303	강인수	KISTI	정보시스템연구팀	dbaisk@kisti.re.kr	0000010005
920060301	이승우	KISTI	정보시스템연구팀	swlee@kisti.re.kr	0000010233

IV. 인력정보 검증

수작업으로 인력정보를 구축하는 과정에서 구축자의 실수에 의해 인력 URI 부여 오류가 발생할 가능성이 크다. 이에 본 장에서는 소속정보, 전자우편, 공저자 관계를 참조하여 자동적으로 인력들을 그룹핑하는 방식과 이를 이용하여 수작업 초기에 참고정보로서 도움을 주거나 수작업 구축 이후에 구축 결과와 비교하여 인력정보를 검증하는 방안을 제안한다.

소속정보나 전자우편과 마찬가지로, 공저자 관계는 동명이인을 구별하는데 중요한 단서가 된다. 논문의 공저자들은 서로 간에 학술적으로 밀접한 관계를 맺고 있으며, 공저자 관계를 가진 논문 건수가 많을수록 그 관계 정도는 더 밀접하다고 할 수 있다. 공저자들 사이의 학술적 관계는 어느 정도 지속되는 것이 보통이다. 심지어, 저자가 소속을 바꾼 경우 - 예를 들어 학위 과정을 마치고 같은 분야에서 학술 활동을 계속하는 경우 - 에도 학술적 관계를 계속 유지하는 예를 흔히 볼 수 있다. 즉, 저자의 소속과 전자우편이 같지 않더라도 공저자 관계를 통해 그 저자가 동일인임을 알아내는 것이 어느 정도 가능함을 의미한다. 또한, 공저자 관계에 있는 두 인력이 모두 동명이인 문제를 가지는 경우가 극히 드물기 때문에 그 효용성이 크다고 할 수 있다. 다음은 소속정보, 전자우편, 공저자 관계를 이용한 자동 검증 방법과 각 단계의 예[표 4][표 5]를 보여준다.

동명을 가진 성과정보 각각을 하나의 인력 그룹으로 하여,

1. 소속과 전자우편이 같은 동명들을 그룹핑하여 하나의 인력 그룹으로 취급한다.
2. 공저자를 공유하는 서로 다른 인력 그룹이 존재하는 경우에 두 인력 그룹을 하나의 인력 그룹으로 취급한다.
3. 1과 2 과정을 더 이상의 인력 그룹이 병합되지 않을 때까지 반복한다.

표 4. 소속정보, 전자우편을 이용한 인력 그룹핑

인명	그룹 ID	공저자 관계	소속기관	소속부서	전자우편
김종원	ID1	한상우; 이동훈; 김종원;	광주과학기술원	정보통신공학과	jongwon@kjist.ac.kr
김종원	ID2	김종원; 최영복;	동명정보대학교	정보통신공학과	kjwyes22@empal.com
김종원	ID3	강민수; 김종원; 이원철; 신요안;	숭실대학교	정보통신전자공학부	
김종원	ID4	김종원; 박성희; 김대영;	한국전자통신연구원		jongwkim@etri.re.kr
김종원	ID1	권영우; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@kjist.ac.kr
김종원	ID1	홍기원; 김양근; 최덕재; 김종원; 박주원; 정종렬;	광주과학기술원		jongwon@kjist.ac.kr

김종원	ID11	김종원; 이장욱; 이원철; 유명식; 신요안;	숭실대학교	정보통신전자공학부	ocosjwt@amcs.ssu.ac.kr
김종원	ID5	이승주; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr
김종원	ID6	송경민; 손주현; 황지수; 김종대; 이용업; 류문호; 김종원;	㈜바이오메드랩		
김종원	ID10	손주현; 황지수; 송경민; 김종대; 이선우; 류문호; 김종원;	㈜바이오메드랩		
김종원	ID12	이장욱; 김종원; 신요안;	숭실대학교	정보통신전자공학부	
김종원	ID7	이철호; 최정용; 권영우; 김종원; 신지태; 김재곤;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@gist.ac.kr
김종원	ID8	한상우; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@nm.gist.ac.kr
김종원	ID5	윤하영; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr
김종원	ID5	박상훈; 이승주; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr
김종원	ID9	황구연; 김재윤; 이동훈; 김종원; 이진영; 주성순;	광주과학기술원	정보통신과	jonwon@netmedia.gist.ac.kr
김종원	ID5	김남곤; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr

표 5. 공저자 관계를 이용한 인력 그룹핑(회색 배경은 공저자 관계를 통해 해소된 인력)

인명	그룹 ID	공저자 관계	소속기관	소속부서	전자우편
김종원	ID1	한상우; 이동훈; 김종원;	광주과학기술원	정보통신공학과	jongwon@kjist.ac.kr
김종원	ID2	김종원; 최영복;	동명정보대학교	정보통신공학과	kjwyes22@empal.com
김종원	ID3	강민수; 김종원; 이원철; 신요안;	숭실대학교	정보통신전자공학부	
김종원	ID4	김종원; 박성희; 김대영;	한국전자통신연구원		jongwkim@etri.re.kr
김종원	ID1	권영우; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@kjist.ac.kr
김종원	ID1	홍기원; 김양근; 최덕재; 김종원; 박주원; 정종렬;	광주과학기술원		jongwon@kjist.ac.kr
김종원	ID3	김종원; 이장욱; 이원철; 유명식; 신요안;	숭실대학교	정보통신전자공학부	ocosjwt@amcs.ssu.ac.kr
김종원	ID5	이승주; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr
김종원	ID6	송경민; 손주현; 황지수; 김종대; 이용업; 류문호; 김종원;	㈜바이오메드랩		
김종원	ID6	손주현; 황지수; 송경민; 김종대; 이선우; 류문호; 김종원;	㈜바이오메드랩		
김종원	ID3	이장욱; 김종원; 신요안;	숭실대학교	정보통신전자공학부	

김종원	ID1	이철호; 최정용; 권영우; 김종원; 신지태; 김재곤;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@gist.ac.kr
김종원	ID1	한상우; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@nm.gist.ac.kr
김종원	ID5	윤하영; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr
김종원	ID5	박상훈; 이승주; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr
김종원	ID1	황구연; 김재윤; 이동훈; 김종원; 이진영; 주성순;	광주과학기술원	정보통신과	jonwon@netmedia.gist.ac.kr
김종원	ID5	김남곤; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr

상기 과정을 거쳐 자동 생성된 인력 그룹정보는 구축 초기에 참고하거나 구축 후 인력 URI 부여 오류를 발견하기 위한 용도로 이용된다. [표 6]은 구축자가 수작업으로 과학기술인력 종합정보시스템을 통해 확인하여 최종 구축한 인력정보의 예를 보여준다.[6]

표 6. [표 5]의 인력정보 검증 결과를 참고하여 구축한 인력정보의 예(회색 배경은 과학기술인력 종합정보시스템에 등록된 인력)

인명	그룹 ID	공저자 관계	소속기관	소속부서	전자우편
김종원	6410138462	한상우; 이동훈; 김종원;	광주과학기술원	정보통신공학과	jongwon@kjist.ac.kr
김종원	0000006015	김종원; 최영복;	동명정보대학교	정보통신공학과	kjwyes22@empal.com
김종원	0000006014	강민수; 김종원; 이원철; 신요안;	숭실대학교	정보통신전자공학부	
김종원	0000006013	김종원; 박성희; 김대영;	한국전자통신연구원		jongwkim@etri.re.kr
김종원	6410138462	권영우; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@kjist.ac.kr
김종원	6410138462	홍기원; 김양근; 최덕재; 김종원; 박주원; 정종렬;	광주과학기술원		jongwon@kjist.ac.kr
김종원	0000006014	김종원; 이장욱; 이원철; 유명식; 신요안;	숭실대학교	정보통신전자공학부	ocosjwt@amcs.ssu.ac.kr
김종원	6410138462	이승주; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@netmedia.gist.ac.kr
김종원	0000006012	송경민; 손주현; 황지수; 김종대; 이용업; 류문호; 김종원;	(주)바이오메드랩		
김종원	0000006012	손주현; 황지수; 송경민; 김종대; 이선우; 류문호; 김종원;	(주)바이오메드랩		
김종원	0000006014	이장욱; 김종원; 신요안;	숭실대학교	정보통신전자공학부	
김종원	6410138462	이철호; 최정용; 권영우; 김종원; 신지태; 김재곤;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@gist.ac.kr
김종원	6410138462	한상우; 김종원;	광주과학기술원	정보통신공학과 네트워크미디어연구실	jongwon@nm.gist.ac.kr

6 본 연구를 통해 최종적으로 구축한 인력정보는 초기에 자동으로 인력 그룹핑을 수행한 후 결과를 참고하고, 수작업 구축 후 다시 한 번 인력정보 검증을 거쳐 자동 인력 그룹핑과 불일치된 비교결과를 확인한 것이다.

김종원	6410138462	윤하영; 김종원;	광주과학기술원	정보통신 공학과 네트워크 미디어연 구실	jongwon@n etmedia.gist. ac.kr
김종원	6410138462	박상훈; 이승주; 김종원;	광주과학기술원	정보통신 공학과 네트워크 미디어연 구실	jongwon@n etmedia.gist. ac.kr
김종원	6410138462	황구연; 김재윤; 이동훈; 김종원; 이진영; 주성순;	광주과학기술원	정보통신과	jonwon@net media.gist.ac .kr
김종원	6410138462	김남곤; 김종원;	광주과학기술원	정보통신 공학과 네트워크 미디어연 구실	jongwon@n etmedia.gist. ac.kr

본 논문에서 제시한 인력정보 검증 방안과 인력정보 구축 방안은 그 과정에 있어 유사함(예를 들어, 공저자나 소속정보를 이용하는 부분 등)을 가지고 있다. 그렇지만 수작업 구축 과정과 유사한 방식으로 기계적인 판단을 함으로써 구축자의 실수가 발생한 경우나 서지정보의 일부가 누락된 경우를 재확인할 수 있다는 측면에서 구축 작업의 정확성을 높일 수 있으며, 또한 미리 인력 그룹핑을 자동으로 수행함으로써 수작업의 효율성을 높일 수 있다.

V. 온톨로지 관계(Property) 활용

인력정보 구축 작업을 수행할 때 완전하게 동명이인 문제를 해소할 수 없는 경우가 발생할 수 있다. 예를 들어, 한 인력이 소속을 변경하여 논문을 작성한 경우에는 공저자, 전자우편, '소속+연도' 모두에 대해 불일치할 수 있다. 또한, 국가과학기술인력 종합정보시스템에서도 해당 인력의 경력이 제대로 갱신되지 않을 수 있다. 이런 경우들에 대해 현재로서는 다른 인력으로 간주할 수밖에 없는데, 추후 경력정보가 갱신되거나, 인력 홈페이지 등의 다른 경로를 통한 확인에 의해 동명이인 문제가 해소될 수 있는 상황을 고려해야 한다.

본 연구에서는 이러한 가능성을 현 시점에서 무리하게 해소하는 대신에 OWL(Web Ontology Language) 기반의 온톨로지 관계(Property)를 이용하여 문제가 해소되는 시점에 처리할 수 있도록 한다. OWL Lite 또는 그 상위 버전(OWL DL, OWL Full)에서 제공하는 동등 관계(Equality Property)들 중에서 'sameAs' 관계는 서로 다른 이름을 가진 개체(Individual)들이 동일한 자원(Resource)을 지칭한다는 의미를 부여하는데 사용된다[10]. 이 관계를 이용하면 서로 다른 URI들을 가지는 인력들을 동등 관계로 연결함으로써 추론 과정에서 동일 인력으로 처리하는 것이 가능하다. 다음은 OWL 형식으로 인력 URI "0000000169"와 "5910089763"이 동일 인력임을 설정하는 예를 보여준다.

```
<Person rdf:ID="0000000169">
<owl:sameAs rdf:resource="#5910089763" />
</Person>
```

VI. 구축 결과

표 7. 과학기술 문헌 구축 결과(외부 기반정보)[7]

학회명	연도	학술대회명	건수 (사용/수집)
한국정보과학회	2002	춘계학술대회	555/591
		추계학술대회	757/768
	2003	춘계학술대회	774/794
		추계학술대회	870/874
	2004	춘계학술대회	682/729
대한전자공학회	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
한국통신학회	2003	추계학술대회	481/481
	2004	추계학술대회	352/352
	2005	춘계학술대회	481/511
		통신정보합동학술대회	215/215
한국인터넷정보학회	2005	추계학술대회	170/170
	2006	춘계학술대회	80/80

한국컴퓨터산업교육학회	2005	추계학술대회	8/8
한국과학기술정보인프라 워크숍	2005	학술대회	141/141
한국멀티미디어학회	2005	추계학술대회	227/227
	2006	춘계학술대회	154/154
총합			9,484/9,770

외부 인력정보 구축을 위해 수집하고 사용한 과학기술 문헌 종류 및 크기는 [표 7]과 같다. 사용 건수와 수집 건수가 다른 것은 중복 논문들과 원본을 얻을 수 없는 논문들(PDF, HWP, MS-OFFICE 등 서식 문서들 중 텍스트를 추출할 수 없거나 원본이 존재하지 않는 논문들)을 배제했기 때문이다.

중복을 포함하여 전체 저자 출현 회수는 24,500건(논문 당 평균 공저자수는 약 2.58명)이며, 한 논문의 최대 공저자 수는 17명이다. 동명이인을 포함하여 동일 이름 저자의 최대 출현 회수는 55회이며, 2,169명은 1회만 출현한 저자들이다.

표 8. URI 기반 인력정보 구축 결과[8]

내용	건수(비율)
외부 인력정보	
성과물 수(= 동명이인 수)	22,331
Unique 성과물 수	8,624
동일 인명 수	5,146
Unique 인력 수	8,487
국가과학기술인력 종합정보시스템 등록 인력 수 (Unique 인력 수 - 신규 URI 부여 인력 수)	1,332 (15.7%)
내부 인력정보	
Unique 인력 수	486
국가과학기술인력 종합정보시스템 등록 인력 수 (Unique 인력 수 - 신규 URI 부여 인력 수)	98 (20.2%)

외부 인력정보 구축은 동명이인 문제 해소를 중점적으로 다루기 위해 2번 이상 성과물에서 나타난 인명들만을 대상으로 하였다. II장과 III장에서 정의한 인력 URI 획득 및 생성 방법을 통해 구축한 결과는 [표 8]과 같다.

외부 인력정보에 대한 검증을 위해 성능 평가 수단으로서 Rand Index를 사용하였다[7]. Rand Index는 모든 인력 쌍들에 대해 정답 집합(수작업 구축 결과)과 대상 집합(자동 인력정보 검증 결과) 간의 동일 그룹으로서 일치하는 수(#Agreement)와 불일치하는 수(#Disagreement)를 이용한다. 특정 인력 쌍이 정답 집합과 대상 집합 모두에서 같은 그룹에 속하거나 모두 다른 그룹에 속하면 일치로 계산하고 그 외의 경우에는 불일치로 계산한다.

$$P = \frac{\#agreement}{\#agreement + \#disagreement}$$

성능 P는 상기 식과 같이 전체('일치+불일치')에 대한 일치 비율로 그룹핑의 정확도를 측정한다. #Under-clustering은 불일치 결과 중 정답 집합에서 동일 인력으로 설정되어 있지만 대상 집합에서 비동일 인력으로 설정되어 있는 경우를 말하며, #Over-clustering은 불일치 결과 중 정답 집합에서 비동일 인력으로 설정되어 있지만 대상 집합에서 동일 인력으로 설정되어 있는 경우를 말한다.

표 9. 외부 인력정보 검증 결과(수작업에 의한 인력정보 구축결과와 자동 인력정보 검증 결과 간 비교)

항목	결과
외부 성과정보 1차분	#Agreement: 56,052 #Disagreement: 4,190 Performance(P): 56,052/60,242 = 0.930447 #Over-clustering: 541 #Under-clustering: 3,649
외부 성과정보 2차분	#Agreement: 3,507 #Disagreement: 230 Performance(P): 3,507/3,737 = 0.938453 #Over-clustering: 1 #Under-clustering: 229

7 본 연구에서 사용한 과학기술 문헌은 3차에 걸쳐 수집 및 가공되었다. 1차(한국정보과학회, 대한전자공학회, 한국 HCI 학회, 한국정보처리학회)와 2차(한국통신학회)는 학술대회논문집 CD를 입수하여 가공되었으며, 3차(한국인터넷정보학회, 한국컴퓨터산업교육학회, 한국과학기술정보인프라워크숍, 한국멀티미디어학회)는 한국과학기술정보연구원에서 서비스하는 학술논문관리 자동화시스템(Article Contribution Management System: ACOMS)에 등록된 문헌들을 입수하여 가공되었다.

8 건수는 외부 인력정보와 내부 인력정보가 중복되는 경우는 고려하지 않은 수치이다.

외부 성과정보 3차분	#Agreement: 1,126 #Disagreement: 378 Performance(P): 1,126/1,504 = 0.748670 #Over-clustering: 60 #Under-clustering: 318
외부 성과정보 123차분(총합)	#Agreement: 72,788 #Disagreement: 5,940 Performance(P): 72,788/78,728 = 0.924550 #Over-clustering: 1,511 #Under-clustering: 4,429

[표 9]를 살펴보면 외부 성과정보 3차분을 제외하고는 모두 90% 이상의 높은 성능을 보인다. 외부 성과정보 3차분의 경우 전자우편은 대부분 들어 있지만 소속정보가 빠진 경우가 많으며 검증 대상 문헌 건수가 작음으로 인해 동명이인 출현 빈도가 상대적으로 낮아서 공저자 관계를 활용하기 어렵기 때문으로 파악되며, 이는 [표 10]에서도 확인할 수 있다.

표 10. 인력정보 검증 결과 분석

문헌 유형	소속정보 부재 비율	동명이인 2회 출현 비율
외부 성과정보 123차분	12.67%	18.29%
외부 성과정보 3차분	44.36%	43.02%

정답 집합과 대상 집합 간 불일치 수(#Disagreement)에 대해 수작업으로 다시 한 번 검증한 결과 103명(전체 대비 0.46%)이 잘못 구축된 것으로 판명되어 이를 수정하였다. 결국 인력정보를 자동으로 검증하는 방안은 인력정보 구축 초기나 이후에 효율적인 인력정보 구축이 될 수 있도록 도움을 주나, 문헌의 크기에 따라 그 효율성은 영향을 받는다.

VII. 결 론

본 연구는 실제 과학기술 문헌을 대상으로 온톨로지에서 기본적으로 요구되는 URI 기반 인력정보를 구축하는 방안과 결과를 설명하였다. 이러한 URI 기반 인력정보는 의미 기반 정보검색, 추론 서비스와 같이 문자열만으로는 해결할 수 없는 고급 응용 서비스 분야에서 중요한 기반정보로서의 역할을 할 것으로 기대한다. 또한, 연구 성과에 대한 정확한 통계를 구하는데 있어서도 핵심이 되는 것이 인력정보이므로 그 활용 가능성은 상당히 크다고 할 수 있다. 향후 연구에서는 개인 홈페이지를 포함한 웹 페이지 등을 참조하여 자동적인 수집 및 정보추출을 통해 방대하고 정확한 인력정보 구축이 이루어질 수 있도록 할 예정이다.

참 고 문 헌

[1] 구회관, 정한민, 강인수, 성원경, 이승준, 심빈구, "국가 과학기술 R&D 기반정보 온톨로지 구축을 위한 URI 관리 및 서비스 시스템 구현", 한국컴퓨터종합학술대회, pp.217-219, 2006.

[2] 신동구, 김재수, 윤정모, 권이남, 전성진, 정택영, "식별체계 간 연계시스템 구축에 관한 연구", 한국정보과학회 추계학술대회, pp.895-897, 2005.

[3] 이상환, 신동구, 김재수, 최진영, 정택영, "식별체계 기반의 과학기술분야 전자원문 연계시스템 설계 및 구현", 한국정보과학회 춘계학술대회, pp.415-417, 2004.

[4] 장대근, 지수영, 오원근, 김의정, "논문 첫 페이지 영상의 논리적 구조형성을 위한 제목, 저자, 요약 영역의 추출", 한국정보처리학회 추계학술대회, pp.1357-1361, 1998.

[5] 정한민, 강인수, 구회관, 이승우, 성원경, "URI 서버에 기반한 국가 R&D 기반정보 온톨로지 설계 및 구현", 정보관리연구, 제37권, 제2호, pp.109-126, 2006.

[6] 정한민, 이승우, 강인수, 성원경, "온톨로지 구축 지원을 위한 과학기술 문헌으로부터의 인력정보 구축", 한국콘텐츠학회 춘계종합학술대회, pp.223-226, 2006.

[7] H. Alani, S. Dasmahapatra, N. Gibbins, H. Glaser, S. Harris, Y. Kalfoglou, K. O'Hara, and N. Shadbolt, "Managing Reference: Ensuring

Referential Integrity of Ontologies for the Semantic Web," In Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), pp.317-334, 2002.

[8] http://www.hrst.or.kr/hrst/index.jsp

[9] http://www.aktors.org/technologies/csaktivespace/

[10] http://www.w3.org/TR/owl-features/

저 자 소 개

정 한 민(Han-Min Jung) 정회원

- 1992년 2월 : 포항공과대학교 전자계산학과(공학사)
- 1994년 2월 : 포항공과대학교 전자계산학과(공학석사)
- 2003년 8월 : 포항공과대학교 컴퓨터공학과(공학박사)
- 1994년~2000년 : 한국전자통신연구원 선임연구원
- 2000년~2004년 : (주)다이퀘스트 연구소장/기술이사
- 2004년~현재 : 한국과학기술정보연구원 선임연구원
- 2004년~현재 : 과학기술연합대학원대학교 겸임교수

<관심분야> : 자연어처리, 시맨틱 웹, 정보추출, 정보검색

이 승 우(Seung-Woo Lee) 정회원

- 1997년 2월 : 경북대학교 컴퓨터공학과(공학사)
- 1999년 2월 : 포항공과대학교 컴퓨터공학과(공학석사)
- 1999년~2000년 : 포항공과대학교 정보통신연구소 연구원
- 2005년 8월 : 포항공과대학교 컴퓨터공학과(공학박사)
- 2005년~2006년 : 대구가톨릭대학교 컴퓨터교육과 강의전담교원
- 2006년~현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 자연어처리, 시맨틱 웹, 정보검색

강 인 수(In-Su Kang) 정회원

- 1995년 2월 : 경북대학교 컴퓨터공학과(공학사)
- 1999년 2월 : 포항공과대학교 컴퓨터공학과(공학석사)
- 2006년 2월 : 포항공과대학교 컴퓨터공학과(공학박사)
- 1995년~1997년 : (주)포스데이타
- 1999년~2001년 : 포항공과대학교 학술정보원
- 2006년~현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 자연어처리, 시맨틱 웹, 정보검색

성 원 경(Won-Kyung Sung) 정회원

- 1987년 2월 : 연세대학교 불어불문학과(학사)
- 1989년 2월 : 연세대학교 불어불문학과(석사)
- 1996년 12월 : 프랑스 파리7대학교 언어학과(박사)
- 1997년~1998년 : 한국전자통신연구원 Post-doc
- 1998년~2001년 : L&H Korea(주) 책임연구원
- 2001년~2003년 : (주)보이스텍 연구개발본부장/상무이사
- 2004년~현재 : 한국과학기술정보연구원 정보시스템연구팀장/책임연구원
- 2004년~현재 : 과학기술연합대학원대학교 겸임교수

<관심분야> : 자연어처리, 시맨틱 웹