한국기록관리학회지, 17(4), 2017, pp. 177-200

기록정보 LOD 구축을 위한 의미 상호연결 자동화 실험 연구*

An Experimental Study on the Automatic Interlinking of Meaning for the LOD Construction of Record Information

하 승 록(Seung-rok Ha)**
안 대 진(Dae-Jin An)***
임 진 희(Jin-hee Yim)****

목 차

〈초 록〉

빅데이터, 인공지능 등 신기술 환경에서 LOD는 기록정보자원을 내외부의 다양한 데이터들과 연결되도록 할 것이다. 이러한 연결의 중심에는 상호연결(Interlinking) 기술이 존재하며, 상호연결된 LOD는 기록정보 개방을 데이터 개방(Open Data)의 최상위 단계로 실현할 것이다. 지속적으로 증가하는 기록의 양을 감안하면, LOD 구축 시 상호연결 알고리즘을 통한 자동화는 필수적이다. 이에 본 연구는 기록정보가 외부 데이터와 상호연결되는 구조와 상호연결 시 고려해야 할 기록정보의 특성을 분석하였다. 또한 국가기록원 CAMS 데이터의 샘플을 수집하여 기록정보 LOD를 구축한 뒤, 기록물 메타데이터의 인물정보를 DBPedia와 자동으로 상호연결하는 테스트베드를 진행하였다. 이를 통해 상호연결 자동화 프로세스를 확인하고, 자동화 기술의 성능과 정확도를 확인하였다. 그리고 테스트베드를 통해 얻은 시사점을 통해 기록정보 LOD 상호연결 과정의 고려사항을 파악하였다.

주제어: 기록정보서비스, 시맨틱 웹, 기록 온톨로지, Record in Context, 상호연결 자동화, SILK, 기록 접근점

〈ABSTRACT〉

In a new technological environment such as big data and AI, LOD will link record information resources with various data from both inside and outside. At the heart of this connection is the interlinking technology, and interlinked LOD will realize the opening of record information as the highest level of open data. Given the ever-increasing amount of records, automation through interlinking algorithms is essential in building LODs. Therefore, this paper analyzed the structure of record information interlinking with the external data and characteristics of the record information to be considered when interconnecting. After collecting samples from the CAMS data of the National Archives, we constructed a record information's LOD. After that, we conducted a test bed that automatically interlinks the personal information of the record metadata with DBPedia. This confirms the automatic interlinking process and the performance and accuracy of the automation technology. Through the implications of the testbed, we have identified the considerations of the record information resources of the LOD interlinking process.

Keywords: Records Information Service, Semantic Web, archival ontology, Record in Context, automatic interlinking, SILK, archival access point

---

* 본 연구는 2017년 국가기록원 R&D사업 '차세대 기록관리 모델 재설계 연구'의 일환으로 수행된 연구임.
** ㈜아카이브랩 연구원(gktmdfhr@gmail.com) (제1저자)
*** 명지대학교 기록정보과학전문대학원 박사과정, (주)아카이브랩 대표(daejin@archivelab.co.kr) (공동저자)
**** 정보인권연구소 연구위원(yimjhkr@empas.com) (교신저자)
■ 접수일: 2017년 10월 31일 ■ 최초심사일: 2017년 11월 12일 ■ 게재확정일: 2017년 11월 16일
■ 한국기록관리학회지 17(4), 177-200, 2017. 〈http://dx.doi.org/10.14404/JKSARM.2017.17.4.177〉

1. 머리말

1.1 연구배경 및 필요성

빅데이터, 인공지능, 사물인터넷, 클라우드 컴퓨팅과 같은 신기술이 발달한 새로운 환경 속에서 기록정보의 양은 더욱 급증할 것이다. 데이터활용법과 데이터기반법(안) 등으로 인해 데이터 기반의 행정이 요구됨에 따라 공공기관의 업무도 빅데이터 분석을 활용하는 쪽으로 변화하고 있다. 이제 빅데이터 속에서 원하는 정보는 찾는 것뿐만 아니라 이 데이터들을 연계하고 분석하여 새로운 가치를 만들어 내는 것이 중요해졌다. 웹의 창시자인 팀 버너스 리(Berners-Lee, T)는 정보의 유의미한 연결과 공유가 커다란 힘이 된다고 말했고,[1] 새로운 가치를 창출하는 데이터의 연결을 위해 데이터를 개방하고 공유하는 움직임은 이미 전 세계적인 추세가 되었다. 그리고 이러한 움직임의 중심에는 기계가 데이터의 의미와 맥락을 이해할 수 있도록 도와주는 시맨틱 웹(Semantic Web) 기술이 존재한다.

시맨틱 웹이란 별도의 웹이 아니라, 잘 정의된 의미를 정보에 부여함으로써 기계와 사람이 더 효과적으로 협력할 수 있는 현재 웹의 확장을 말한다(Berners-Lee, Hendler, & Lassila, 2001). 시맨틱 웹의 데이터는 컴퓨터가 이해할 수 있는 기계가독형 언어로 되어있기 때문에 컴퓨터가 이용자의 요구에 적합한 결과를 찾아주는 의미기반 검색이나 인공지능의 학습데이터로 활용된다.[2] 링크드 오픈 데이터 Linked Open Data: LOD)는 이러한 시맨틱 웹을 구현하기 위한 기술적인 방법이자 접근법이다. 이미 의료, 전자정부, 전자상거래, 도서관, 박물관 등 다양한 분야의 데이터들이 LOD로 구축되어 웹에 발행되었다(김용겸, 2014).

LOD는 차세대 기술 환경의 주요한 기반이 될 것이다. 이는 LOD가 서로 다른 형태의 기계나 사물, 서비스들이 데이터의 맥락을 이해하고 추론하는 것을 기술적으로 돕기 때문이다. LOD를 구축할 때는 데이터에 의미를 부여하고 관계를 정의하면서, 데이터의 의미를 외부자원(데이터세트)의 의미와 연결하는 작업을 진행하는데, 이러한 의미의 연결을 상호연결(Interlinking)이라 한다. 즉 상호연결은 내부의 LOD(RDF 데이터)를 외부의 LOD와 연결함으로써 개념에 대한 공통된 의미를 공유하고, 이 개념을 나타낼 수 있는 설명・표현・속성을 웹에서 공유하며, 개념에 대한 의미를 명세화하는 작업이다. 예를 들어 '배'라는 데이터가 있다면 인간은 직관적으로 이것이 과일 배인지, 운송수단 배인지, 사람 몸의 배인지 알 수 있다. 하지만 기계는 이를 알지 못하기 때문에 외부 데이터세트에서 '배(Boat)'라는 의미를 가진 데이터와 의미를 연결해줌으로써 기계가 의미를 이해할 수 있도록 하는 것이다. 상호연결은 데이터가 고립된 섬(Data Silo)에서 벗어나 전 세계적인 웹의 공간으로 연결되고, 응용 프로그램이 데이터 소스를 검색할 수 있게 해주는 접착제 역할을 한다(Heath & Bizer, 2011). 또한

1) 팀 버너스 리의 TED 강연(https://www.ted.com/talks/tim_berners_lee_on_the_next_web) 내용 중 일부이다.
2) IBM의 인공지능 기술인 왓슨(Watson)은 시맨틱 웹을 구현하기 위한 기술인 링크드 오픈 데이터(Linked Open Data: LOD)를 지식 소스로 활용하여 질의응답 프로그램을 개발하였다.

상호연결은 서로 다른 형태의 데이터에 대해서 의미를 통합하는 역할을 한다. 이러한 상호연결은 시맨틱 웹이 추구하는 데이터의 개방과 공유라는 목적을 달성하기 위해서 중요한 역할을 한다. 하지만 계속적으로 증가하는 기록의 양을 감안하면, 거대한 양의 LOD 데이터를 인간이 일일이 확인하고 대조하여 상호연결하는 것은 물리적으로 불가능하다. 따라서 상호연결 품질의 하락을 감수[3]하더라도 상호연결의 자동화는 반드시 필요하다.

1.2 선행연구

링크드 데이터와 온톨로지에 관한 연구로, 하승록, 임진희, 이해영(2017)은 오픈소스 도구를 이용한 기록정보 LOD 구축 절차와 방법을 살펴보고, 테스트베드를 수행하여 기록관 LOD 구축을 위한 필요요건을 제시하였다. 박옥남(2012)은 국가기록원의 전거데이터세트를 링크드 데이터로 구축하는 연구를 진행하였다. 연구를 통해 기록물의 개방과 공유를 위해 기록물 온톨로지를 설계하고, 링크드 데이터로 구축하는 방법을 제시하였다. 이유빈, 이해영(2017)은 온톨로지 기반의 기록물 검색 시스템을 위한 인터페이스를 제안하였다. 연구를 통해 시맨틱 검색 방법, 카테고리 및 패싯, 검색 결과 탐색 및 재검색, 상세정보 제시 등의 측면에서 특징을 분석하고 이용자 평가를 진행하였다. 윤소영(2013)은 링크드 데이터 구축 사례를 바탕으로 공공데이터 활용을 위한 국가 연계체계를 제안하였다. 연구를 통해 국가 연계체계사업인 공공DB 피디아 구축과정에서 도출된 문제를 파악하고 기존의 국가DB 연계체계 구축과정을 참고하여 그 해결 방안을 제시하였다. 이성숙, 박지영, 이혜원(2017)은 링크드 데이터에서 인물 정보의 식별 및 연계 범위 확장에 관한 연구를 진행하였다. 연구를 통해 인물과 관련된 별도의 전거 정보 구축을 통해 서지데이터 검색의 접근점을 확장하는 방안을 제시하였다. 새로운 기록물 기술표준에 관한 연구로, 박지영(2016)은 차세대 기록물 기술표준에 관한 연구 진행하여 국제적인 기록물 기술표준의 개정 동향을 살펴보고, 이와 같은 동향이 국내의 기록물 기술표준의 발전에 주는 시사점을 도출하였다. 또한 박지영(2017)은 ISAD(G)에서 RiC-CM으로의 전환에 관한 연구를 통해 새로운 기록기술 표준인 Record in Context의 개념과 특징에 대해 분석하였다. 상호연결(Interlinking)과 상호연결 자동화 방법에 대한 저서와 연구로, Heath와 Bizer(2011)는 링크드 데이터에 관한 저서에서 상호연결의 정의와 유형에 대해 제시하였다. Singh(2011)는 LOD 상호연결의 방법과 도구를 분석하고, SILK 프레임워크를 분석하였다. Schaible와 Mayr(2012)는 SILK를 활용하여 LOD의 상호연결을 진행해보고 시사점을 정리하였다. Isele(2013)은 개체(Entity) 매칭을 위한 연결 규칙에 대한 연구에서 SILK Framework의 개요와 연결 규칙, 비교연산자 등에 대해 분석하였다. Auer, Bryl, & Tramp(2014)는 유럽 위원회(European Commission)에서 진행된 LOD2 프로젝트 연구 결과를 제시하며, 링크드 데이터를 상호연결하기 위한 자동

3) 인간은 직관적으로 두 개체의 의미가 동일함을 인지할 수 있기 때문에 LOD의 상호연결(Interlinking) 작업은 인간이 직접 했을 때 그 품질이 가장 높다고 말한다.

화 도구로 SILK Link Discovery Framework를 제시하였다.

1.3 연구목적 및 방법

본 연구는 기록정보에 LOD가 도입되는 의미를 분석하고, 기록정보의 상호연결과 상호연결 자동화 필요성에 대해 분석하고자 한다. 또한 기록정보 LOD 상호연결의 구조와 고려사항을 분석하고, 테스트베드를 통해 기록정보의 상호연결 자동화 프로세스를 확인하고, 자동화 기술의 성능과 정확도를 확인하고자 한다. 그리고 테스트베드 과정에서 얻은 시사점을 통해 기록정보 LOD 상호연결의 고려사항을 파악해 보고자 한다.

본 연구는 문헌연구와 실험연구로 진행되었다. 문헌연구는 LOD의 요소 기술에 대한 논문이나 조사보고서 등을 조사하여 기록정보에 시맨틱 웹과 LOD가 도입되는 것이 가지는 의미에 대해서 분석하고, 기록정보의 의미 상호연결 필요성과, 상호연결 자동화 필요성에 대해서 분석하였다. 테스트베드 대상 기록정보는 국가기록원 홈페이지에서 서비스하고 있는 기록물 시리즈로 선정하였다. 선정한 기록물 시리즈의 CAMS 데이터를 요청하고, 제공받은 데이터를 활용하여 LOD를 구축한 후에는 오픈소스 소프트웨어인 SILK Workbench를 활용하여 외부자원과 연계하는 테스트베드를 수행하였다. 그리고 테스트베드 결과를 통해 기록정보의 상호연결 및 자동화에 대한 시사점을 도출하였다.

2. 기록정보 LOD 구축의 과제

2.1 시맨틱 웹과 LOD 도입의 의미

시맨틱 웹은 현재 웹에서 사용하는 기술(HTML, HTTP, URI,[4] Hypertext)을 활용하면서, 정보자원에 잘 정의된 개념(의미)이나 구조화된 정보를 부여함으로써 현재 웹의 확장된 모습이라 할 수 있다. 이러한 시맨틱 웹은 정보의 검색・통합・자동화 측면에서 현재 웹과는 다른 모습을 보인다. 먼저 기계가 정보자원의 의미를 파악할 수 있는 시맨틱 웹은 의미 기반 검색을 통해 복잡한 질의요구에 대응하여 현재의 용어 매치 형태의 검색보다 적합한 결과를 가져올 수 있다. 그리고 RDF(Resource Description Framework)[5] 기반의 데이터 표현을 통해 서로 다른 어플리케이션간의 자유로운 데이터 상호교환 및 통합을 이룰 수 있다. 마지막으로 정보자원에 의미(Semantic)와 기술정보를 연관시킴으로써 기계가 데이터의 의미를 파악하고 자동으로 관계를 추론하여 웹 서비스를 제공할 수 있도록 업무를 자동화할 수 있다.

시맨틱 웹 환경을 만들기 위한 핵심적인 기

4) URI(Uniform Resource Identifier)는 웹에서 하나의 개체를 식별하기 위한 식별자로 활용한다.

5) RDF란 URI에 의해 식별가능한 모든 웹의 자원을 표현하기 위한 프레임워크이다. RDF에서는 모든 데이터를 주어(표현하는 대상), 술어(자원의 속성), 목적어(속성의 값)로 표현한다. RDF에서 주어부, 술어부, 목적부로 구성되는 서술문을 트리플(triple)이라고 부른다. 추상적인 RDF그래프모델은 컴퓨터가 직접 이해할 수 없으므로 컴퓨터가 이해하고 처리할 수 있는 기계적인 언어인 XML기반의 RDF Syntax를 사용하여 표현하며, RDF Syntax는 RDF/XML, N-Triple, N3, Turtle 등이 있다.

술로는 웹 자원을 표현하는 방식인 RDF, 온톨로지를 기술하는 언어인 OWL,[6] 구조화된 데이터를 발행하고 연결하는 데이터 구조인 LOD (Linked Open Data)[7] 등이 있다. 기록정보를 시맨틱 웹에 맞게 발행한다는 것은 RDF, OWL과 같은 온톨로지 언어를 활용하여, 정해진 원칙과 표준에 따라 기록정보를 기계가 이해할 수 있는 기계가독형 언어로 발행하는 것을 의미한다.

시맨틱 웹의 대표적인 기술인 링크드 오픈 데이터(LOD) 형태로 기록정보를 발행한다면 아카이브는 다음과 같은 이점들을 가질 수 있다. 첫째, LOD 기술을 활용한 시맨틱 검색은 기존의 키워드 검색이나 시소러스 연계 검색과 다르게 이용자의 복잡한 검색 의도를 반영할 수 있기 때문에 높은 재현율과 정확률을 보장하여 이용자의 검색 만족도 향상을 기대할 수 있다. 둘째, LOD는 다양한 표준포맷(RDF, SKOS,[8] OWL 등)을 기반으로 데이터를 공유하기 때문에 기록정보가 박물관, 도서관, 혹은 전혀 다른 성격의 정보자원과 연결될 수 있다. 외부자원과 상호연결된 LOD는 관계된 데이터의 계속되는 연결을 통해 예상하지 못한 잠재적 지식을 발견(Serendipity)할 수 있는 장점이 있다. 따라서 다른 분야의 정보자원 데이터세트에서 기록정보를 연결할 수 있는 환경이 조성된다면, 기록정보의 활용이 더욱 활발해지고 다양해지는 것을 기대할 수 있다. 셋째, 기록정보를 소장한 아카이브 입장에서 다른 데이터세트의 데이터를 URI를 통해 재사용함으로써 데이터의 중복 구축을 방지하고, 중복데이터 재생산 과정을 축소할 수 있다. 이는 아카이브가 기록정보만의 고유한 데이터 생산에 주력하여 데이터의 품질 향상을 가져오는 것을 기대할 수 있다.[9] 넷째, 이용자가 프로그래밍 방식(기계가독형)으로 아카이브의 데이터에 쉽게 접근하고, 한 번에 많은 양의 데이터를 가져다가 사용할 수 있도록 개방할 수 있다. 이는 기록정보에 대한 일반이용자 및 외부기관의 접근과 활용도를 높여 기록정보의 가치를 제고하는 것을 기대할 수 있다.

---

6) OWL(Web Ontology Language)은 RDF/RDFS의 한계 때문에 만들어졌으며 다른 온톨로지 마크업 언어들보다 표현력이나 추론능력이 뛰어나다는 평가를 받고 가장 널리 사용된다. OWL은 온톨로지 헤더(Header), 클래스(Class), 속성(Property), 객체(individual) 등 4개의 구성요소를 가진다. 온톨로지 헤더(Header)는 OWL 온톨로지 문서 전체의 정보를 기술하는 요소이고, 클래스(Class)는 비슷한 속성을 지니고 있어 하나의 군으로 모아지는 객체(individual)들을 의미하고, 속성(Property)은 객체간의 관계 혹은 객체와 데이터 값의 관계를 표현하는 어휘를 의미한다.

7) 연결된 데이터(Linked Data)와 열린 데이터(Open Data)의 개념이 합쳐진 개념이다. Linked Data란 기술적인 개념으로, 웹페이지가 서로 연결된 것처럼 데이터들끼리 다양한 관계에 의해 연결되어 있는 형태를 말한다. Open Data란 문화적인 개념으로, URI를 이용해서 누구나 데이터를 접근할 수 있으며 데이터에 대한 정보를 제공받을 수 있는 형태를 말한다.

8) SKOS(Simple Knowledge Organization System)는 W3C에서 시맨틱 웹 기반의 지식 구조화 프레임워크 표준으로 개발한 것으로, SKOS를 이용하면 지식 표현 체계를 시맨틱 웹으로 표현하는 것이 가능하다(박옥남, 2012).

9) 국가기록원이 개발한 '기록물 생산기관 변천정보'와 같은 전거데이터를 LOD로 구축하여 발행한다면, 다른 기록관의 LOD를 구축할 때, 생산기관에 대한 LOD를 구축할 필요 없이 국가기록원이 발행한 LOD를 가져와서 쓸 수 있다. 이는 한국형 공공정보데이터 LOD Cloud 구축에 큰 기여를 하고, 새로운 부가가치 창출 기여할 수 있을 것이다.

2.2 기록정보의 의미 상호연결 필요성

데이터는 21세기에 들어서 '제 2의 원유(Data is the new oil)'라 불리며, 데이터를 생산・활용・응용・분석하기 위한 기술적・학문적 노력이 지속되고 있다(한국정보화진흥원 지식자원활용부, 2015). 2015년 European Data Portal 보고서에 따르면, 공공데이터의 개방에 따른 시장규모가 3,250억 유로(€)만큼 성장하였고, 2만 5천개의 일자리가 생겨났으며, 17억 유로(€) 만큼의 공공행정 비용을 절약했다(European Commission, 2015). 기록정보의 상호연결은 기록정보가 더 이상 아카이브의 데이터베이스 안에만 고립되어 있는 것이 아니라 외부 데이터와의 연계를 통해 새로운 부가가치를 창출하기 위해 반드시 필요한 작업이다. 특히 기록정보는 다른 정보들과는 다르게 체계적으로 관리되어 왔고, 다양한 기술요소를 통해 많은 정보가 기술되어 왔다. 따라서 기록정보가 LOD로 구축되어 웹에 발행된다면 많은 정보를 개방하고 공유하여 큰 부가가치를 창출할 수 있을 것으로 기대된다.

상호연결은 〈표 1〉과 같이 3가지 종류가 있다. LOD에서 상호연결을 진행하는 가장 대표적인 방법은 표목 역할을 하는 데이터세트와 내부데이터를 연결하는 것이다. 예를 들어 국립중앙도서관의 한국 국가서지 LOD의 경우, 미의회도서관의 주제명/저자명 표목표와 상호연결 진행하여 서지에 대한 데이터를 연계하였다. 또한 한국 국가서지 LOD는 국내의 LOD 사이트(KDATA, NDSL, RISS, 생물정보, 국립공원관리공단, 특허청)와의 데이터 연계를 통해 서지정보에 대한 상호연결을 활발하게 구축하고 있다. 기록정보도 기록의 생산기관, 기능, 인물(단체명), 장소, 주제, 기록물 형식 등의 접근점을 통해 외부 데이터세트와 상호연결이 가능할 것이다.

상호연결은 기록정보 LOD가 새로운 애플리케이션 제작을 지원하는 것으로 활용될 수 있다. 예를 들어 기록관이 사진기록에 대한 장소 정보를 DBPedia 혹은 주소데이터 LOD와 같은 외부 데이터세트와 상호연결한다면, 기관이 직접 장소에 대한 정보를 구축하지 않더라도 상호연결한 외부 데이터세트에서 장소의 위도와 경도에 대한 정보를 가져와서, 그것을 통해 〈그림 1〉과 같이 장소의 지리정보를 활용한 시각화 서비스를 제공할 수 있다.

〈표 1〉 상호연결(Interlinking)의 종류(Heath & Bizer, 2011)

종류	내용
관련링크 (Relationship Links)	주석(annotation)과 같은 데이터와 관련된 정보를 연결시켜주는 링크
동질 링크 (Identity Links)	객체와 동일하거나 추상적인 개념을 식별하기 위해서 dc나 skos, foaf와 같은 다른 전거 어휘를 활용하여, 온톨로지에서 사용한 속성(혹은 클래스)과 동일한 의미를 가진 어휘를 연결시켜주는 링크
어휘링크 (Vocabulary Links)	용어의 의미에 대한 정의를 나타내기 위해 owl:sameAs와 같은 온톨로지 속성을 사용하여, 동일한 의미를 가진 객체를 연결하는 링크

〈그림 1〉 기록정보 LOD를 활용한 응용 서비스 예시

2.3 개념의 의미 상호연결 자동화 필요성

LOD의 상호연결은 인간이 수동으로 연결하는 방식(Schema Dependent)과 시스템이 자동으로 연결하는 방식(Schema Independent)이 있다. 수동연결 방식은 RDF 술어(Predicate)의 의미에 관한 지식이 필요하므로[10] 인간이 수동으로 작업을 수행할 수밖에 없다. 따라서 수동연결 방식은 상호연결 품질의 우수성을 기대할 수 있으나 대량의 트리플 데이터(LOD의 RDF 데이터)의 상호연결 작업을 처리할 수 없다. 반면에 자동연결 방식은 데이터 구조(스키마)에 대한 인간의 지식이 필요로 하지 않으므로 자동화가 가능하다(이경욱, 2015). 따라서 LOD로 구축하고자 하는 기록정보의 양이 방대하다면 상호연결 작업의 자동화는 반드시 필요하다. 상호연결의 자동화 방법은 키 기반 방법과 유사성 기반 방법이 있다. 키 기반 방법은 데이터 집합이 ISBN,[11] GTIN, ISIN 등과 같이 표준 식별자를 포함하는 경우 역함수 속성을 이용하여 동일 개체 여부를 판단하는 형태인데, 특정한 경우에만 적용가능하다는 한계가 있다. 도서의 경우 ISBN과 ISSN[12]과 같은 국제표준식별자를 사용하기에 이 방법을 사용하기에 적합하지만, 기록정보의 경우 국제표준식별자를 사용하지 않기 때문에 이 방법을 사용하기는 적합하지 않다. 유사성 기반의 상호연결 자동화 방법으로는 SILK, LIMES, SERIMI, SLINT, AgreeMaker 등이 있다. 유사성 기반의 기본적인 원리는 사용자가 지정하거나 많이 언급된

10) A 데이터세트의 술어와 B 데이터세트의 술어(predicate)가 같은 의미라는 것을 알아야 한다.
11) ISBN(International Standard Book Number)은 국제적으로 책에 붙이는 고유한 식별자이다.
12) ISSN(International Standard Serial Number)은 국제 표준 간행물 번호이다.

술어를 비교키로 선정하고 그에 따른 목적어 값을 비교하여, RDF 트리플의 주어 개체들이 어느 정도 일치하는지를 알고리즘을 통해 파악하여, 유사도가 높으면 상호연결하는 방식이다. 이는 대량의 트리플 데이터를 처리할 순 있지만, 단순하게 목적어 값을 비교하는 것만으로는 상호연결의 품질이 떨어질 수 있다는 단점이 있다. 상호연결 자동화 방법을 택하게 된다면, 어느 데이터세트와의 상호연결이 자관 LOD 데이터의 활용을 효과적으로 높일지에 대한 분석이 필요하다.

3. LOD 상호연결의 구조와 기록정보의 특성

3.1 사전의 종류와 역할

LOD에서 정보의 의미를 명확하게 정의하는 것은 외부에 이미 존재하는 사전과의 연결을 통해 이뤄진다. 의미의 명세화를 위해 사전과 연결한다는 것은 어휘(Vocabulary)를 재사용하는 것과 표목 데이터세트와의 상호연결(Interlinking)을 진행하는 것을 뜻한다.

RDF 형태로 이루어진 LOD는 주어, 술어, 목적어의 트리플 그래프로 자원을 기술한다. 그러나 이것은 추상적인 데이터 모델일 뿐이지, 그 객체와 관계에 대한 특정한 영역의 의미를 표현하지는 않는다(Heath & Bizer, 2011). 따라서 LOD는 객체(클래스)와 관계(속성)의 의미를 명확하게 하기 위해서 이미 구축된 어휘를 재사용한다. LOD 구축에 있어서 광범위하게 사용되는 어휘는 Dublin Core Metadata Initiative (DCMI),[13] Friend of a Friend(FOAF),[14] SKOS[15] 등이 있다. 어휘를 재사용 한다는 것은 예를 들어 데이터의 제목에 대한 관계를 표현할 때, 새로 어휘를 만드는 것보다 더블린 코어의 “제목(title)”에 해당하는 어휘를 가져와서 재사용하는 방식이다. 이러한 표준 어휘의 재사용은 외부 응용 프로그램을 통해 자관의 데이터를 검색하기 쉽게 만들어, 데이터의 활용도를 높일 수 있다. 광범위하게 사용되는 어휘를 재사용하는 것 이외에도 특정 분야의 프레임워크를 재사용하여 어휘의 의미를 좀 더 세분화하는 방법도 있다. 예를 들어 TV 방송 분야의 온톨로지인 Programmes Ontology,[16] 도서관 분야의 데이터 모델인 BIBFRAME,[17] 문화유산 분야의 온톨로지인 CIDOC-CRM(Conceptual Reference Model)[18] 등이 있다. 기록학 분야에서도 최근 온톨로지로 활용이 가능한 새로운 기록물 기술표준인 Record in Context(RiC)를 발표하였다(ICA/EGAD, 2016a). 이렇게 특정 분야(영역)의 온톨로지를 재사용하여 어휘를 표현하는 것은 LOD의 성격을 분명히 정의하며 데이터의 의미를 특정 분야의 의미로 표현해준다. LOD를 구축할 때 기본적으로 특

13) Dublin Core Metadata Initiative. http://dublincore.org/documents/dcmi-terms/
14) Friend of a Friend. http://xmlns.com/foaf/0.1/
15) SKOS. http://www.w3.org/2004/02/skos/
16) Programmes Ontology. https://www.bbc.co.uk/ontologies/po
17) BIBFRAME. https://www.loc.gov/bibframe/
18) CIDOC-CRM. http://www.cidoc-crm.org/

정 영역의 어휘를 사용하면서도 온톨로지 설계 단계에서 표준 어휘와 동일한 관계임을 설정한다면(예를 들어 RiC:Name=dc:Title) 표준 어휘의 범용성까지 획득할 수 있다.

상호연결은 데이터가 외부로 연결되는 수단이자 동시에 데이터의 의미를 명확하게 정의하며, 추가적인 기술(Description)없이 외부에서 데이터의 기술을 획득할 수 있는 방법이다. 따라서 데이터의 의미를 명확하게 정의하는 표목 역할의 데이터세트에 대한 중요성이 부각된다. 대표적인 표목 역할을 하는 LOD 데이터세트는 〈표 2〉와 같다.

특히나 DBPedia는 2007년 베를린 자유대학과 라이프치히 대학의 공동 연구로 시작된 프로젝트로, 위키피디아로부터 구조화된 정보를 자동으로 추출하여 RDF로 변환하고 SPARQL을 통해 질의할 수 있도록 LOD로 공개된 데이터세트이다. DBPedia는 125개 언어에 대해 30억개의 정보(RDF 트리플)로 구성되어 있으며 다양한 분야의 정보를 가지고 있기 때문에 활용성이 높고, LOD 분야에서 중요한 역할을 수행하고 있다. DBPedia는 많은 LOD 데이터세트들이 개념의 의미를 명확하게 정의하기 위하여 상호연결하고 있으며, 약 5천만 개의 다른 LOD 데이터세트와 연결되어 있다. 국내 LOD 사이트에서도 대다수가 DBPedia와 상호연결을 진행하였다.

3.2 의미 상호연결을 위한 메타데이터 추출

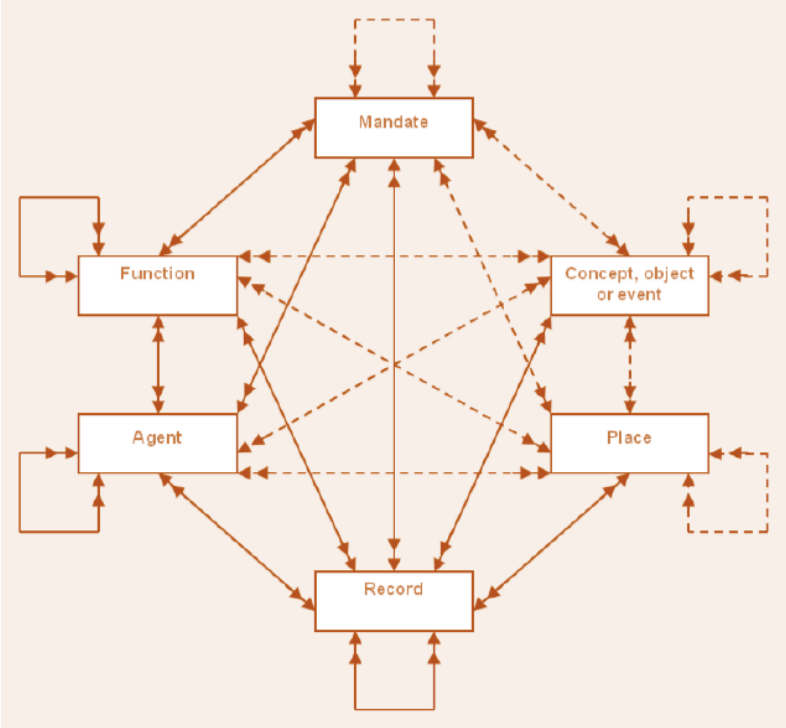
기록정보를 LOD로 구축하면서 상호연결 작업을 진행할 때는 기록정보의 기술요소 중에서 외부 데이터세트와 연결될 수 있는 메타데이터 항목을 추출해야 한다. 상호연결할 메타데이터 항목을 추출할 때는 외부에서 아카이브의 기록정보로 접근할 수 있는 연결점을 고려해야 한다. 보통 일반 이용자들이 키워드(주제어)검색을 선호하는 것을 고려할 때, 검색의 키워드로 사용되는 요소가 상호연결하기 좋은 메타데이터 항목일 것이다. 앞선 선행연구들에서는 상호연결 하기 좋은 접근점으로 이름, 장르, 주제, 직업 등을 분석하였고(Gracy, 2015), 특히 인물정보는 상호연결의 핵심적인 요소로 활용된다고 밝혔다(이성숙, 박지영, 이혜원, 2017). 이와 같은 요소들을 확인해봤을 때, 기록관리의 다중개체 모형에서 제시하는 전거 개체들이 상호연결을 위한 접근점으로 활용될 수 있을 것이다. 새로운 기록기술 표준인 Record in Context(RiC)

〈표 2〉 대표적인 표목 데이터세트

데이터세트	설명
DBPedia	위키피디아의 데이터를 LOD로 구축한 데이터세트이다. LOD 분야에서 개념의 의미를 명세화하기 위해 가장 널리 사용되는 데이터 허브이다.
Geonames	1000만개 이상의 지리적 정보와 장소 정보, 위도, 경도, 인구 정보 및 우편번호 등의 정보를 무료로 제공하는 데이터세트이다.
MeSH	의학분야의 주제명표목(시소러스)을 RDF문서로 제작한 데이터세트이다.
LC Linked Data Service	미국 의회도서관에서 공표한 표준과 어휘에 대해서 제공함으로써 미국 의회도서관 링크드 데이터 서비스에서 제공하는 모든 어휘에 대해서 명확한 사용 방침을 제공한다.

에서는 기존 기술표준(ISAAR(CPF), ISDF, ISDIAH)의 전거개체(생산자, 기능, 소장기관)들 뿐만 아니라 문서형식, 날짜, 장소, 개념/대상 등도 그 중요성을 인정하여 기존에는 속성이었던 것을 개체(Entity)로 승격시켰다(박지영, 2017). 따라서 RiC에서 제시하는 14개의 개체 중에서 행위자(생산기관, 인물), 업무기능, 기록물 형식, 날짜, 장소, 개념/대상(주제) 등이 기록정보를 외부 데이터와 상호연결하기 좋은 포인트(접근점)가 될 것이다.

의미를 상호연결하기 위한 메타데이터 항목의 추출이 끝나면, 해당 메타데이터를 어떤 외부 LOD 데이터세트와 상호연결할 것인가에 대한 분석이 필요하다. 기록정보의 의미를 명확히 정의하는 상호연결도 필요하지만(예를 들어 표목 데이터세트와의 상호연결), 기록정보 LOD 데이터의 활용을 효과적으로 높일 수 있는 LOD 데이터세트가 무엇이 있는지 분석이 필요하다. 기록정보 LOD가 외부 LOD와 상호연결되면 이용자는 의미 추적을 통해 다른 데이터세트로 이동하여, 의도하지 않은 잠재된 지식을 획득할 수 있다. 예를 들어 사진기록의 촬영장소가 "남산"이었고, 그 장소를 DBPedia의 "남산"과 상호연결한다면, 아카이브가 따로 "남산"에 대한 정보를 기술하지 않아도 이용자는 의미의 상호연결을 통해 DBPedia가 가지고 있는 "남산"에 대한 초록, 위도, 경도, 사진 등 다양한 정보들을 획득할 수 있다.



〈그림 2〉 Record in Context의 개체 관계도(ICA/EGAD, 2016b)

3.3 기록정보 키워드 추출 단계의 고려사항

LOD에서 데이터가 연결되는 접근점은 인물, 장소, 주제 등 주로 데이터의 실질적인 내용과 관련된 값이다. 하지만 기록정보의 기술(Description)은 주로 계열(series) 등 상위계층의 기술정보에 집중되어 있다. 집합적 기술 등 기록학의 전통적 원칙들은 개별 기록의 내용보다는 전체적인 구조나 맥락을 설명하는 데 집중해 왔기 때문이다. 국가기록원 CAMS의 기록정보 역시 철이나 건 단위보다는 계열 등 상위계층의 메타데이터가 풍부하다. 따라서 앞의 메타데이터 추출과정과 같이 메타데이터를 분석하여 LOD 연결을 위한 키워드를 추출하거나 본문 내용으로부터 LOD 연결점이 될 만한 키워드를 추출하여 건 메타데이터로 확보하는 작업이 필요하다.

하지만 현실적으로 방대한 양의 기록물 건에 대해서 아키비스트가 모든 내용 정보를 기술할 수는 없다. 따라서 인공지능과 같은 차세대 기술을 활용하여 방대한 양의 기록물의 내용 메타데이터를 확보하는 방안을 모색해야 한다. 인공지능을 활용한 텍스트 분석, 이미지 인식, 음성인식 등의 기술은 상용화 단계로 성숙되었다.[19] 현재 인공지능 기술은 키워드를 추출하여 키워드가 인물인지 장소인지 개념인지를 판단하고, 문서의 내용을 파악하여 자동분류 하는 수준까지 이르렀고, 구글과 IBM, 아마존과 같은 기업에서는 클라우드 인공지능 API 서비스를 제공하고 있다. 그러나 아카이브가 자체적으로 기술을 개발하고 이를 실행할 전문 인력을 배치하는 것은 현실적으로 불가능하다. 따라서 인공지능 관련 기업의 상용 서비스를 이용하여 기록물의 내용 메타데이터를 확보하는 방안을 모색해야 할 것이다. 이렇게 인공지능 기술을 활용하여 기록정보의 내용 메타데이터를 확보하고, 기록정보를 LOD로 구축하기 위해서는 다음과 같은 전처리 작업이 선행되어야 한다. 첫째, 텍스트 추출 전략이 필요하다. 종이기록과 전자기록 등 다양한 유형별로 별도의 디지털화 및 텍스트 추출 전략이 수립되어야 한다. 예를 들어 종이기록의 경우 수기로 작성되었거나 한자나 일어가 포함되어 OCR 기술을 적용해도 인식하기가 쉽지 않은 편이다. 전자기록 중에서도 HWP 등 독자적인 비트스트림을 채택한 문서 포맷보다는 XML 형태로 제목, 본문, 저자 등의 항목이 구조화된 ODF 포맷이 키워드 추출에는 훨씬 유리할 것이다. 이처럼 기록의 다양한 포맷별로 별도의 방법론을 적용해야 한다. 둘째, 한글처리 기술 또한 요구된다. 기록의 내용으로부터 키워드를 추출한 이후 형태소 분석을 통해 인물이나 장소, 사건 등을 별도의 클래스로 인식하여 입수할 수 있어야 한다. 텍스트 분석 및 자연어 처리, 인공지능 기술을 활용하기 위한 방안을 모색해야 한다.

3.4 상호연결 자동화 프로세스

본 연구는 유사성 기반의 상호연결 자동화 방법 중에서 SILK Link Discovery Framework

19) 인공지능 기술 분야에서 한글 처리에 대한 기술성숙도는 아직 미흡하나, 구글과 IBM에서 2017년 하반기에 한글 인식이 가능한 인공지능 플랫폼 서비스를 오픈하였고, ETRI의 엑소브레인과 같이 한글 처리가 능숙한 인공지능 기술이 발전하고 있다.

를 활용하였다. SILK 프레임워크는 사용자가 설정한 연결 규칙 문서인 SILK-LSL(Link Specification Language)에 따라 상이한 데이터의 의미를 연결하여 RDF 링크를 생성한다(Auer et al., 2012). 연결규칙은 선언적 언어[20]를 사용하여 표현되며, 데이터 항목이 상호연결되기 위해 따라야 하는 조건을 정의한다. 이러한 연결 규칙은 유사성 기반의 방법으로 비교 측정의 조건을 정의한다(Auer, Bryl, & Tramp, 2014). SILK Link Discovery Framework는 SILK Single Machine,[21] SILK MapReduce,[22] SILK Server[23] 등 3가지의 명령행(Command-line) 기반 애플리케이션과 웹 기반 어플리케이션이 있다(Euzenat et al., 2011). 본 연구는 JAVA 기반 웹 애플리케이션인 SILK Workbench를 활용하였다. SILK Workbench는 사용자가 연결 규칙을 쉽게 작성하고 편집할 수 있는 그래픽 기반의 편집기능을 제공하고, 연결 규칙의 유효성을 자동으로 검증하는 등 이용자 친화적인 인터페이스를 제공하고 오픈소스 소프트웨어이기 때문에 누구나 무료로 활용 가능하다. SILK의 LOD 상호연결 프로세스를 간단히 도식화하면 〈그림 3〉과 같다.

SILK가 서로 다른 두 데이터세트의 데이터를 비교하여 의미를 연결하는 것은 연결 규칙(Linkage Rules) 설정을 통해 이뤄진다. 연결규칙 설정에서는 먼저 서로 비교할 데이터세트에서 SPARQL 검색이나 RDF 파일 업로드를 통해 데이터를 가져온다. 둘째, 서로 비교할 메타데이터 항목(RDF의 속성)을 RDF Property Path 설정을 통해 지정한다. 예를 들어 A 데이터세트의 "hasName"과 B 데이터세트의 "Title"을 선정하는 것이다. 셋째, 선정한 메타데이터의 값(RDF 속성의 값)을 서로 비교하기 위하여 변환(Transformations)을 통해 속성의 값을 정제한다. 예를 들어 대소문자를 통일한다거나 특수문자를 제거하거나, URI에서 값을 정제하는 등의 행위가 있다. 넷째, 데이터의 속성 값을 비교하기 위하여 다양한 유사성 메트릭(similarity metric) 기법을 사용한다. 유사성 측정 방법은

〈그림 3〉 SILK의 상호연결 프로세스

20) 선언형 프로그래밍 언어를 말하며, 연결 규칙은 XML 태그를 사용하여 XML 문서로 작성된다.
21) 동일한 시스템에 있거나 SPARQL 검색을 통해 접근 가능한 데이터세트를 사용하여 단일 시스템에서 RDF 링크를 생성하는 애플리케이션이다. 본 연구에서 활용한 SILK Workbench가 이것에 해당한다.
22) 여러 시스템의 클러스터를 사용하여 데이터 집합간에 RDF 링크를 생성하는 등 Hadoop(빅데이터 프레임워크)을 기반으로 빅데이터 환경에서 SILK를 사용할 수 있는 애플리케이션이다.
23) LOD를 사용하는 웹 환경의 응용 프로그램 내에서 HTTP API를 통해 상호연결 값을 제공하는 애플리케이션이다.

문자 기반, 토큰(Token) 기반, 특수 영역 기반으로 유사성을 비교한다. SILK에서 제공하는 대표적인 유사성 메트릭의 종류는 〈표 3〉과 같다. 다섯째, 여러 개의 유사성 측정 결과 값을 종합(Aggregations)하여 최종적인 유사도 평가를 진행한다. 종합 기능(Functions)은 최소값(Minimum), 최대값(Maximum), 가중 평균(Weighted Average) 등이 대표적이다.

4. 상호연결 자동화 테스트베드

4.1 실험의 필요성 및 목표

차세대 웹에서는 기계가 이해하고 추론할 수 있는 형태의 데이터를 선호하며, 기존의 데이터베이스 데이터들을 링크드 오픈 데이터(LOD) 형태로 구축하여 개방하고 공유하는 추세가 확산되고 있다. 이러한 LOD의 개방 및 공유의 핵심은 다양한 분야의 데이터와 상호연결하는 것에 있다. 그러나 국내 LOD 구축사업 결과물을 살펴보면 상호연결이 상당히 제한적으로 진행된 것을 볼 수 있다. 따라서 본 연구는 알고리즘을 활용한 자동화 도구를 통해 기록정보 LOD의 상호연결 작업을 수행하는 테스트베드를 진행하였다. 이러한 테스트 베드를 통해 본 연구는 기록정보 LOD의 데이터 객체를 외부 LOD와 자동으로 상호연결하기 위한 프로세스를 확인하고, 상호연결 자동화 기술의 성능과 정확도 확인하고자 한다. 그리고 테스트베드를 통해 얻은 결과를 통해 기록정보 LOD 구축 및 상호연결 자동화 과정에 대한 시사점을 도출하고자 한다.

4.2 실험 시스템 환경 및 프로세스

샘플 기록정보 LOD 구축은 TopQuadrant사의 Topbraid Composer Free Edition[24]을 사용하였다. 상호연결 자동화 방법인 SILK Workbench는 Github를 통해 오픈소스로 공개되고 있는 Release 2.7.0-RC1 버전[25]을 사용하였다. SILK Workbench는 JAVA기반의 웹 애플리케이션이므로

〈표 3〉 유사성 메트릭의 대표적인 종류(Isele, R., 2011)

유사성 메트릭	사용
Levenshtein Distance	문자(string)의 유사성 측정 시 사용
Jaro/Jaro-Winkler	문자(string)의 유사성 측정 시 사용
Korean Translit Distance	한글 문자(string)의 유사성 측정 시 사용
Jaccard coefficient	토큰(token)의 유사성 측정 시 사용
Cosine	토큰(token)의 유사성 측정 시 사용
Geographic Distance	지리적 정보의 유사성 측정 시 사용
Date/Time	날짜/시간 정보의 유사성 측정 시 사용
Numbers	숫자의 유사성 측정 시 사용

24) Topbraid Composer Free Edition. https://www.topquadrant.com/downloads/topbraid-composer-install/
25) SILK Workbench 2.7.0-RC1. https://github.com/silk-framework/silk/tree/release-2.7.0-RC1

JDK(Java SE Development Kit) 8 이상을 설치해야 한다.[26] 또한 도스 커맨드 창(cmd)에서 명령어를 실행하기 위한 Simple Build Tool(sbt)[27]을 설치해야 한다. 시스템 환경이 구축되면, 도스 명령어 창에서 [sbt "project workbench" run]을 실행하면 SILK Workbench가 구동되고, 브라우저에서 http://localhost:9000으로 접속하면 애플리케이션 사용이 가능해진다.

테스트베드의 프로세스는 다음과 같다. 첫째, 테스트베드를 위하여 국가기록원 홈페이지 검색을 통해 샘플 기록 컬렉션을 선정하고, 국가기록원에 선정한 기록물의 중앙영구기록관리시스템(CAMS) 메타데이터를 요청하였다. 둘째, 현재 관리되고 있는 CAMS 메타데이터 항목이 새로운 기록기술 표준인 Record in Context(RiC)와 어떻게 매핑되는지 분석하고, 제공 받은 CAMS 메타데이터에서 기록정보 LOD 상호연결을 위한 접근점으로 인물정보를 추출하였다. 셋째, TopBraid Composer라는 온톨로지 및 LOD 구축 소프트웨어를 사용하여 기록정보 LOD를 구축하였다. 넷째, SILK Workbench를 활용하여 샘플 기록정보 LOD의 인물정보를 한국 DBPedia 내에 있는 인물정보와 상호연결하였다. 이 때 상호연결 테스트 조건을 다르게 설정하여 2번의 테스트를 진행하였다. 마지막으로 상호연결 자동화 테스트베드를 통해 얻은 결과를 분석하여 시사점을 정리하였다.

4.3 샘플 기록정보 LOD

본 연구에서는 테스트베드를 진행하기 위하여 샘플 대상 기록정보를 선정하였다. 국가기록원 홈페이지 검색을 통해서 기록정보 LOD를 구축하였을 때 유의미한 시사점이 있을 법한 컬렉션으로 국제스포츠대회[28]와 국무회의기록[29]을 선정하였다. 국제스포츠대회 기록 컬렉션은 다양한 인물, 장소, 주제가 기록에 담겨 상호연결 할 수 있는 접근점이 풍부하여 선정하였고, 국무회의기록은 행정문서기록의 특징을 그대로 담고 있는 기록 시리즈이기 때문에 선정하였다.

중앙영구기록관리시스템(CAMS)의 데이터베이스는 600여개 이상의 테이블, 1만개 이상의 칼럼으로 구성되어 있다. 그 중에서 기록물 철과 기록물 건 위주로 메타데이터 항목을 선정하였다. 선정한 메타데이터는 관리기본목록(기록물 철 메타데이터), 관리세부목록(기록물 건 메타데이터), 정부간행물 메타데이터, 시창각물 메타데이터였고, 총 216개의 메타데이터 항목을 선정하여 국가기록원에 데이터를 요청하였다.

제공받은 CAMS 데이터 중에서는 값이 없는 메타데이터 항목도 다수 존재하고 있었고,[30] LOD로 구축하는 데 필요 없는 메타데이터 항목[31]도 다수 존재하고 있었기 때문에, 총 216개의 메타데이터 항목 중에서 필요한 항목의 데이

26) JAVA JDK 8. http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
27) SBT. http://www.scala-sbt.org/
28) 국제스포츠대회. http://theme.archives.go.kr/next/internationalSport/viewMain.do
29) 국무회의기록. http://theme.archives.go.kr/next/cabinet/viewIntro.do
30) 국가기록원에 CAMS 데이터를 요청하기 위하여 메타데이터를 선정할 때, 기록물 철과 건에 대해서 필요 없는 메타데이터 항목을 선정하여 값이 없는 항목도 있었지만, 대다수는 기록물 기술이 되어있지 않아 공란인 경우가 많았다.

터만을 선정하여 〈그림 4〉와 같이 RiC과 매핑 분석하였다.

본 연구에서는 샘플 기록정보 LOD의 내용정보 중에서 인물 정보를 DBPedia와 상호연결하고자 한다. 제공받은 CAMS 데이터(216개 메타데이터 항목, 10,826건 기록물)를 분석하여 기록의 내용 정보 중 인물에 대한 정보가 담겨 있는 28개의 메타데이터 항목을 선정하고, 그 중에서 실제로 데이터가 있는 12개의 메타데이터 항목을 선정하여 총 84명의 인물정보를 추출하였다.

제공받은 CAMS 데이터를 LOD로 구축하기 위하여 TopBraid Composer를 사용하여 LOD로 구축하였다. 다만 테스트베드의 목적이 상호연결에 초점이 맞춰져 있으므로 인물정보에 대해서는 모두 LOD로 구축하였으나, 모든 기록물 건의 기술을 LOD로 구축하지 않았다는 한계가 있다.

4.4 의미 상호연결 자동화 결과

테스트베드는 대상 데이터세트에서 SPARQL 검색을 통해 데이터를 가져오는 조건(쿼리)을 다르게 설정하여 두 번의 테스트를 진행하였다. 테스트 쿼리 1은 한국 DBPedia의 인물정보 중 국적(nationality), 나라(country), 출생지(birthPlace), 사망지(deathPlace)가 "대한민국"인 사람을 불러오도록 설정하였다. 테스트 쿼리 2는 한국 DBPedia의 인물정보 중 초록(abstract)에 '대한민국'이라는 문자가 들어간 사람을 불러오도록 설정하였다. 이렇게 상

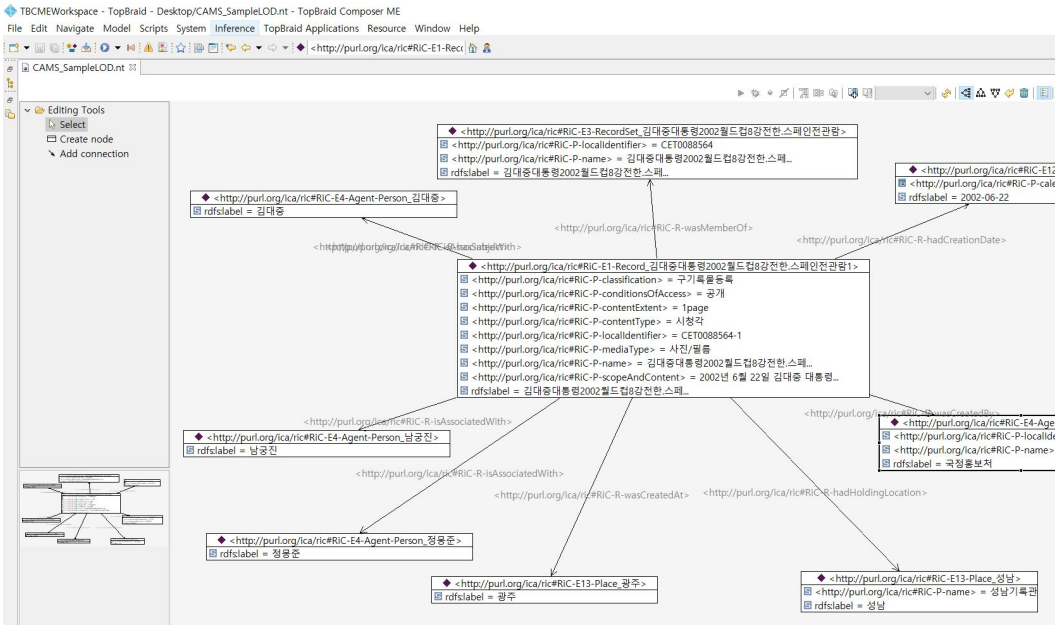
CAMS 테이블의 컬럼명		RiC 개체	RiC 속성
처리기관코드	KIKWANCODE	Agent	ric:RiC-P-localIdentifier
구기록물생산기관명	OLDPRODKIKWAN	Agent	ric:RiC-P-name
생산년도	PRODYEAR	Record	ric:RiC-R-hadCreationDate
기록물철제목	BND_TTL	RecordSet	ric:RiC-P-name
문서대장 , B도면, C사진/필름, D녹음/동영상	DOCTYPE	Record	ric:RiC-P-mediaType
보존시설 : A=서울, B=부산, C=대전, D=성남	ARCH_PLACE	Place	ric:RiC-R-hadHoldingLocation
공개여부(1:공개, 2:부분공개, 3:비공개)	OPN_YN	Record	ric:RiC-P-conditionOfAccess
정각,04=총독부,05=정부간행물,06=해외기록	ARCAVETYPE	Record	ric:RiC-P-contentType
대통령명	PREZ_NM	Agent	ric:RiC-R-hasSubject
수집년도	INSUYEAR	Date	ric:RiC-R-hadCreationDate
분류체계명	CLSS_NM	Record	ric:RiC-P-Classification
내용요약	CONTEXTMEMO	Record	ric:RiC-P-scopeAndContent
내용요약정보	DISCRIPTION	Agent	ric:RiC-R-isAssociatedWith
생산등록일자	PRODREGDATE	Date	ric:RiC-R-hadCreationDate
제목	JEMOK	Record	ric:RiC-P-name
쪽수	PAGE_1	Record	ric:RiC-P-contentExtent
기록물관리번호	MNGNO_1	RecordSet	ric:RiC-P-localIdentifier
인명	HUMNNAME	Agent	ric:RiC-R-hasSubject
발행일자	PUBDATE	Date	ric:RiC-R-hadCreationDate
간행물제목	TITLE	Record	ric:RiC-P-name
저자명	AUTHOR	Agent	ric:RiC-R-wasCreatedBy
내용(목차)	CONTENTS	Record	ric:RiC-P-scopeAndContent
대통령명	PREZ_NM_1	Agent	ric:RiC-R-hasSubject
사진일자	PIC_DT	Date	ric:RiC-R-hadCreationDate
촬영장소(배경장소)	PIC_PLACE	Place	ric:RiC-R-wasCreatedAt
사진설명	PIC_DESC	Record	ric:RiC-P-scopeAndContent
목차	CR_TYPE + "_" + LINE_NO +	Record	ric:RiC-P-scopeAndContent
인물 키워드	HUMN_SEQ_NO + ": " + KOF	Agent	ric:RiC-R-hasSubject
키워드	KEYWORD_DIV + "_" + KEYW	Record	ric:RiC-P-generalNote

〈그림 4〉 CAMS 메타데이터와 RiC 매핑표 예시

31) 기록물id(Primary Key)와 같이 필요없는 메타데이터도 있었고, 분류체계구분, 구기록물철분류번호와 같이 실제적인 내용을 알 수 없는 데이터베이스의 포린 키(Foreign Key) 등은 LOD 구축에서 제외하였다.

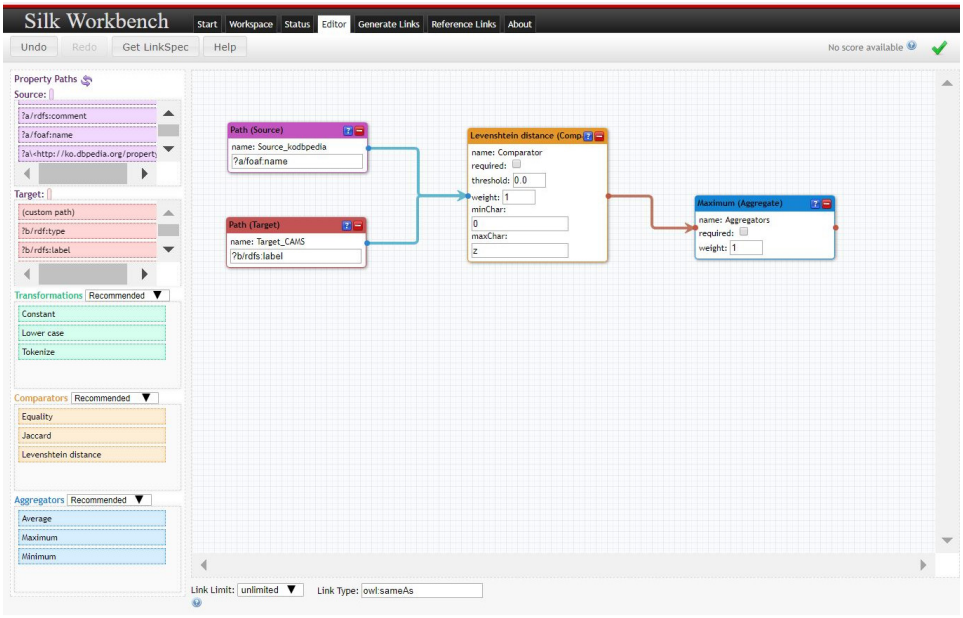
〈표 4〉 인물정보를 포함한 CAMS 메타데이터 항목

테이블	컬럼	예시	
		메타데이터 값	인물 키워드 추출
관리기본목록(철)	기록물철제목	노태우대통령제24회서울올림픽개막식참석	노태우
	대통령명	윤보선	윤보선
	내용요약	1956년 11월 1일 이승만 대통령이, 출국 신고 차 경무대를 방문한 제16회 호주 멜버른 하계 올림픽(11. 22.~12. 8.) 참전 한국선수단(임원 14명, 선수 35명)을 경무대 현관 앞에서 맞이하여, 태극기를 전달하며 국위 선양과 격려의 말을 전한 후 함께 기념촬영을 하였다.	이승만
	내용요약정보	김영삼	김영삼
관리세부목록(건)	제목	김영삼대통령내외분동계올림픽출전선수단오찬시기념품수여	김영삼
	기안자(업무담당자)	이두복	이두복
	인명	박정희, 육영수	박정희 / 육영수
정부간행물	저자명	김종기 외	김종기
시청각세부추가	대통령명	김영삼	김영삼
	사진설명	1982년 12월 6일 정주영 KOC(대한올림픽위원회) 위원장이 김포 공항에서 제9회 뉴델리 아시안 게임 참가 선수단과 귀국 환영 악수.	정주영
주요인물		1:노태우(,); 2:김옥숙(,); 3:김영삼(,)	노태우 / 김옥숙 / 김영삼
키워드정보		일반1:중국인; 일반2:리찬첸; 일반3:동향	리찬첸



〈그림 5〉 구축한 샘플 기록정보 LOD의 기록물 건 개체 관계 시각화

호연결 대상 데이터세트인 DBPedia에서 가져온 인물정보의 이름(foaf:name)을 테스트베드를 통해 구축한 샘플 기록정보 LOD의 인물 이름(rdfs:label)과 값을 레벤슈타인 거리 척도[32]를 통해 비교하여 상호연결이 진행되도록 연결 규칙을 설정하였다.



〈그림 6〉 테스트에서 공통적으로 사용한 연결 규칙(Linkage Rule)

〈표 5〉 SPARQL 질의(Query) 조건

테스트 쿼리 1	SELECT DISTINCT ?a
	WHERE {
	{?a rdf:type foaf:Person ; 〈http://dbpedia.org/ontology/nationality〉 〈http://ko.dbpedia.org/resource/대한민국〉 .} UNION {?a rdf:type foaf:Person ; 〈http://dbpedia.org/ontology/country〉 〈http://ko.dbpedia.org/resource/대한민국〉 .}
	UNION {?a rdf:type foaf:Person ; 〈http://dbpedia.org/ontology/birthPlace〉 〈http://dbpedia.org/resource/대한민국〉 . } UNION {?a rdf:type foaf:Person ; 〈http://dbpedia.org/ontology/deathPlace〉 〈http://dbpedia.org/resource/대한민국〉 .}
	}

테스트 쿼리 2	SELECT DISTINCT ?a
	WHERE {
	?a rdf:type foaf:Person ; 〈http://dbpedia.org/ontology/abstract〉 ?c .
	FILTER REGEX(STR(?c), '대한민국')
	}

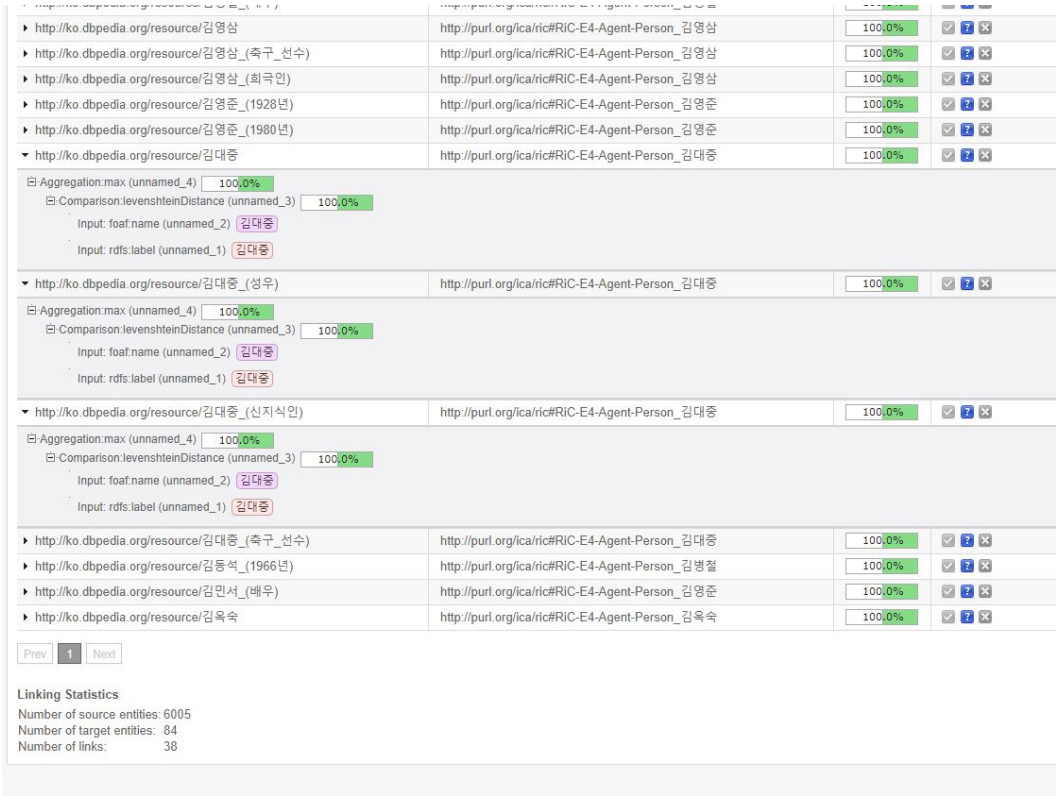
32) 테스트베드에서 단어 수준 유사도 검사 척도로 레벤슈타인 거리 척도와 한글 음역 거리 척도(Korean translit distance)를 고려하였으나, 유의미한 차이가 나타나지 않았다.

테스트베드를 통해 구축한 샘플 기록정보 LOD의 인물정보(84명)를 한국 DBPedia에서 SPARQL 검색해보니 총 48명의 인물정보가 존재하고 있었다. 〈그림 7〉과 같이 SILK Workbench를 통해 상호연결 자동화 테스트를 진행해본 결과, 테스트 쿼리 1을 통해서는 총 808개의 한국 DBPedia 데이터를 불러올 수 있었으며, 48명의 인물정보 중에서 총 9명(18.75%)의 데이터가 상호연결되었다. 그리고 테스트 쿼리 2를 통해서는 총 6,005개의 한국 DBPedia 데이터를 불러올 수 있었으며, 48명의 인물정보 중 38건(79.16%)의 데이터가 상호연결되었다.

다만 테스트 쿼리 2의 경우, SILK는 동명이인을 구분할 수 없기 때문에, 다양한 동명이인들이 샘플 기록정보 LOD의 해당 인물과 모두 연결되었다. 38건의 상호연결 중에서 의미가 동일하게 연결된 상호연결을 분석해보니 48명의 인물정보 중 총 21명(43.75%)의 데이터가 정확하게 상호연결되었다.

4.5 시사점

상호연결 자동화 테스트베드를 수행하여 얻은 결과를 통해 정리한 시사점은 다음과 같다. 이 시사점들은 기록정보 LOD 상호연결 과정의 고려사항을 포함한다.



〈그림 7〉 테스트 쿼리 2를 통한 테스트 결과

첫째, 기록정보 LOD를 외부 데이터세트와 상호연결하기 위한 접근점을 분석하고 외부 데이터세트를 선정해야 한다. 외부에서 기관이 소장한 기록정보로 접근해서 들어올 수 있는 접근점을 분석하는 것은 기록정보의 더 널리 개방되고 더 많이 공유되기 위해서 반드시 필요할 것이다. RiC의 14개의 개체(Entity)는 기록정보가 외부데이터와 상호연결되기 위한 좋은 접근점이 될 것이며, 그 중에서도 어떠한 접근점을 기준으로 상호연결 작업을 진행할 것인지 소장기관의 기록정보를 분석해야할 것이다. 또한 기록정보와 연결되었을 때, 이용자에게 더욱 도움이 될만한 외부 데이터세트로 무엇이 있는지 분석이 필요하다. 유관기관(기록관, 박물관, 도서관 등) 이외에도 전혀 종류가 다른 데이터세트와의 연결도 고려해야 할 것이다.

둘째, 상호연결 대상 데이터세트에 대한 분석이 필요하다. 먼저 대상 데이터세트가 어떤 제약사항이 있는지 파악이 필요하다. 대부분의 데이터세트에서 LOD 데이터를 획득할 수 있는 SPARQL 검색에 제한을 두었다.[33] 이는 LOD 데이터를 저장하는 서버의 부담을 피하기 위해서라고 판단된다. 하지만 앞서 테스트 결과를 봤을 때, 800여건의 데이터를 가져왔을 때와 6,000여건의 데이터를 가져왔을 때의 상호연결 건수의 차이가 극명하다. 이는 대상 데이터세트에서 데이터를 얼마나 가져올 수 있느냐가 상호연결 성공률 향상에 결정적인 역할을 한다고 볼 수 있다. 따라서 대상 데이터세트의 구조에 대한 분석과 이해를 통해 포괄적이면서 정확한 SPARQL 질의문(Query)를 작성해야 한다. 예를 들어 대상 데이터세트가 어떤 어휘(dc, skos 등)를 쓰는지, rdf:type을 무엇으로 쓰고 있는지 등 대상 데이터세트의 온톨로지에 대한 사전 조사가 필요할 것이다. 본 연구에서는 한국 DBPedia에 존재하는 한국인 인물정보를 가져오기 위하여 한국 DBPedia의 온톨로지를 분석하였고, 이를 통해 두 가지 조건(쿼리)을 설계하였다. 그러나 한국인임에도 불구하고 초록(abstract)에 '대한민국'이라는 문자가 없거나, 국적(nationality), 나라(country), 출생지(birthPlace), 사망지(deathPlace)가 "대한민국"이라고 되어 있지 않은 데이터는 검색되지 않은 한계가 있다.

셋째, 기록물의 내용 정보에 대한 기술이나 메타데이터 확보가 필요하다. 일반적으로 기록의 기술은 집합적 기술의 원칙에 따라 개별 기록의 내용보다 전제적인 구조와 맥락을 기술한다. 하지만 LOD 환경과 전자기록환경에서는 데이터의 실질적인 내용이 필요하다. 시청각기록물이 대다수인 "국제스포츠대회" 기록 시리즈는 기록물 건에 대한 내용 정보(인명, 장소 등)가 비교적 자세히 기술되어 있으나, 일반적인 행정문서기록인 "국무회의기록" 기록시리즈는 기록물 철과 건에 대한 내용 정보가 전무하다. 이를 통해서 다른 행정문서 기록 또한 기록의 내용에 대한 기술이 부족할 것으로 짐작할 수 있다. 따라서 기록의 실제적인 내용을 담고 있는 기록물 건에 대한 기술을 보충해야 할 것이다. 하지만 방대한 양의 기록물에 대한 기술작업을 진행하는 것은 많은 사업 예산을 필요로 하고, 그 효용성에 대한 의문으로 일선 기록관

33) DBPedia는 1만건/1회, 국립중앙도서관 LOD는 100건/1회, 서울시 LOD는 100건/1회, 문화융합LOD는 2만건/1회 등으로 SPARQL 검색의 제한이 있다.

리 현장에서의 반대가 극심할 것이기 때문에 실질적으로 진행하기가 지난한 일이다. 하지만 기록정보가 더 이상 기록관 내부에서 잠들어 있지 않고, 다양하게 활용되고 공유되기 위해서는 반드시 기록물의 내용에 대한 기술이 필요하다.[34] 따라서 빅데이터 분석과 인공지능 기술과 같은 차세대 기술이 성숙됨에 따라, 이를 활용하여 기록물 내용 정보를 기술할 수 있는 방법을 모색해야 할 것이다. 기록물의 내용에 대한 정보가 많이 공개될수록 기록정보는 외부의 다양한 데이터와 연결되어, 많은 이용자가 기록물을 보다 쉽게 확인할 수 있을 것이다.

넷째, 기관이 소장한 기록물의 전거정보를 별도로 분석하여 LOD로 구축해야 한다. 테스트베드를 진행하면서 구축한 샘플 기록정보 LOD는 인물정보의 이름만 LOD로 구축되었다. 이는 평면적인 CAMS 메타데이터 값을 다중 개체 모형인 RiC을 활용한 LOD로 구축하면서 생긴 한계이다. 따라서 상호연결 자동화 테스트 결과에서 동명이인 등을 구분하지 못하고 상호연결이 진행된 문제가 있었다. 이렇게 동명이인 등 비교 값은 같지만 의미가 다른 데이터(resource)를 처리하기 위해서는 데이터를 비교하기 위한 다른 맥락정보가 필요하다. 예를 들어 인물정보의 경우에는 한글 이름뿐만 아니라 영어이름, 한자이름, 생년월일, 직업, 지위, 호(號) 등 다른 맥락정보가 있는 경우 동명이인을 구분하여 상호연결 작업을 진행할 수 있을 것이고, 장소정보의 경우에는 위도, 경도, 주소 등의 맥락정보를 통해 의미를 구분할 수 있을 것이다. LOD의 큰 장점이 데이터의 재사용을 통해서 중복된 데이터를 구축하지 않는 것이기는 하지만, 상호연결 자동화 작업의 품질 향상을 위해서는 전거정보에 대한 맥락 구축이 일정 부분 필요할 것이다. 이러한 작업을 하기 위해서는 기관이 소장한 기록정보들에 대해서 어떠한 전거정보가 있는지부터 파악해야 할 것이다.

5. 맺음말

기록정보의 LOD 구축은 기록정보가 아카이브 내부에서 고립되어 소비되는 것(Silo)이 아니라 더 넓은 시맨틱 웹 환경에서 다양하게 활용되고, 새로운 부가가치를 창출할 수 있는 가능성을 열어주는 것이다. 이러한 개방과 공유의 중심에는 LOD의 상호연결 작업이 존재하기 때문에 기록정보의 상호연결의 프로세스와 고려사항에 대한 연구가 필요하다. 본 연구는 기록정보 LOD 구축의 의미와 과제에 대해 분석하고, LOD 상호연결의 구조와 기록정보의 특성에 대해서 분석하며, 기록정보 LOD를 자동으로 상호연결하기 위한 고려사항을 분석하고, 테스트베드를 통해 기록정보 LOD 상호연결의 시사점을 도출하고자 하였다.

이에 본 연구는 기록정보에 LOD가 도입되는 이점으로 시맨틱 검색을 통한 기록정보 검색 강화와 외부 정보자원과의 연계 가능성, 중복 데이터 구축 방지, 기계가독형 개방을 통한 기록정보

34) 기록정보 LOD 구축을 위하여 기록물 기술이 필요하기도 하지만, 우선적으로 기록정보를 웹에서 크롤링해갈 수 있도록 원문을 공개하고, OCR 작업을 진행하며, 구조화된 형태로 기록정보를 제공하는 등의 노력이 필요하다. 이는 인공지능 기술 등 차세대 기술을 적용하기 위해서 기록정보가 기본적으로 갖춰야 할 사전작업이다.

개방의 확대를 제시하였다. 그리고 상호연결의 종류와 의미, 필요성에 대해서 설명하였고, 상호연결 자동화 방법의 종류와 필요성에 대해 설명하였다. 그리고 LOD 상호연결의 구조를 설명하기 위하여 LOD에서 사전의 역할을 하는 데이터세트에 대해서 설명하였고, 기록정보의 기술요소 중에서 어떤 종류의 메타데이터 항목들이 외부의 데이터세트와 의미를 연결시킬 필요가 있는지, 기록정보에서 어떤 전처리 작업을 통해 메타데이터를 추출할 수 있는지 분석하였으며, 상호연결 자동화 방법인 SILK의 프로세스에 대해서 분석하였다. 그리고 국가기록원 CAMS 메타데이터를 수집하여 LOD를 구축하고, 상호연결 자동화 오픈소스 소프트웨어인 SILK Workbench를 활용하여 테스트베드를 진행하였다. 그리고 테스트베드를 통해서 기록정보 LOD를 외부와 자동으로 상호연결하기 위한 프로세스를 확인하고 시사점을 도출하였다. 먼저, 기록정보 LOD를 외부 데이터세트와 상호연결하기 위한 접근점을 분석하고 외부 데이터세트를 선정해야 한다는 것을 확인하였다. 그리고 상호연결 대상 데이터세트의 구조에 대한 이해와 제약사항에 대한 분석이 필요하다는 것을 제시하였다. 그리고 인공지능 등 차세대 기술을 활용한 기록물의 내용 정보에 대한 기술이 필요하다고 제시하였으며, 기관이 소장한 기록물의 전거정보를 별도로 분석하여 LOD로 구축해야 한다고 제시하였다.

본 연구는 기록정보의 특성을 고려한 의미연결 규칙 개발을 위하여 테스트베드를 설계 및 수행하여 기록정보 상호연결 자동화 프로세스를 확인하고, 자동화 기술의 정확도와 성능을 확인하는 것에 초점을 맞췄다. 따라서 테스트베드에 활용한 샘플데이터의 규모가 적고, 인물정보 외에 다른 기록정보 상호연결 접근점에 대한 분석이 부족하며, 상호연결 자동화의 성공률이 높지 않다는 한계가 있다. 따라서 후속 연구에서는 이러한 한계점을 보완하여 기록정보의 연결과 확장을 풍부하게 이룰 수 있는 기록정보의 특성에 대해 확인할 수 있는 연구가 제시되기를 희망한다.

기록정보는 계층성과 집합적 기술 등 다른 분야의 정보들과 다르게 갖는 고유한 특성들 때문에 LOD로 구축하기 어려운 점이 있다. 또한 기록을 생산하고 관리하는 현장에서는 업무 과중 때문에 LOD 구축을 위한 기반작업(예를 들면 기록물 기술 보강)을 진행할 여유가 없다는 의견도 제시된다. 하지만 기록정보는 다른 분야의 정보들과는 다르게 대량의 정보가 다양한 기술요소를 통해 체계적으로 관리되고 보존되어 왔으므로, 그 활용가치에 대한 잠재력이 무궁무진하다. 따라서 기록정보를 빅데이터 분석이나 인공지능 학습 등 외부에서 활용할 수 있도록 개방하는 것은 기존의 기록정보콘텐츠 제작과는 또 다른 기록정보서비스의 영역이 될 것이다. 기록정보 LOD를 구축하는 것은 그러한 서비스의 기반이 될 것이며, 앞으로 기록관은 기록정보를 다양한 정보들과의 연계를 통해 사회 구성원들에게 제공하는 방법에 대해 고민해야 할 것이다.

참 고 문 헌

김용겸 (2014). 시맨틱 웹의 주요 응용사례와 발전방향. 동중앙아시아연구(구 한몽경상연구), 25(3), 65-86.

박옥남 (2012). 기록물 전거통제 기반 Linked data 구축에 대한 연구. 한국비블리아학회지, 23(2), 5-25.

박지영 (2016). 차세대 기록물 기술표준에 관한 연구. 한국기록관리학회지, 16(1), 223-245.

박지영 (2017). ISAD(G)에서 RiC-CM으로의 전환에 관한 연구. 한국기록관리학회지, 17(1), 93-115.

윤소영 (2013). 공공데이터 활용을 위한 링크드 데이터 국가 연계체계 구축에 관한 연구. 정보관리학회지, 30(1), 259-284.

이경욱 (2015). LOD InterLinking. 검색일자: 2017. 10. 19. https://www.slideshare.net/ssuser6e1ce5/interlinking-lod

이성숙, 박지영, 이혜원 (2017). 링크드 데이터에서 인물 정보의 식별 및 연계 범위 확장에 관한 연구. 정보관리학회지, 34(3), 7-21.

이유빈, 이해영 (2017). 온톨로지 기반의 기록물 검색 시스템을 위한 인터페이스 제안. 한국기록관리학회지, 17(1), 217-244.

하승록, 임진희, 이해영 (2017). 오픈소스 도구를 이용한 기록정보 링크드 오픈 데이터 구축 절차 연구. 정보관리학회지, 34(1), 341-371.

한국정보화진흥원 지식자원활용부 (2015). 알기 쉬운 Linked Open Data. 서울: 한국정보화진흥원.

Auer, S., Bryl, V., & Tramp, S. (2014). Linked Open Data - Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project (Vol. 8661). Springer.

Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., ... & Stadler, C. (2012). Managing the life-cycle of linked data with the LOD2 stack. In International semantic Web conference, 1-16. Springer, Berlin, Heidelberg.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.

European Commission (2015). Creating Value throught Open Data: Study on the Impact of Re-use of Public Data Resources. Luxembourg.: Publications Office of the European Union.

Euzenat, J., Abadie, N., Bucher, B., Fan, Z., Khrouf, H., Luger, M., ... & Troncy, R. (2011). Dataset interlinking module. Retrieved October 22, 2017, from https://hal.archives-ouvertes.fr/file/index/docid/793433/filename/datalift-421.pdf

Gracy, K. F. (2015). Archival description and linked data: A preliminary study of opportunities and implementation challenges. Archival Science, 15(3), 239-294.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the Web into a global data space. San Rafael, Calif.: Morgan & Claypool.

ICA/EGAD (2016a). Record In Contexts(RiC) An Archival Desription Draft Standard. 2016 ICA Congress. Retrieved October 20, 2017, from https://www.ica.org/sites/default/files/session-7.8-ica-egad-ric-congress2016.pdf

ICA/EGAD (2016b). Record In Contexts: A Conceptual Model For Archival Description. Consultation Draft v.0.1.

Isele, R. (2011). Link Generation for the Data Web. Retrieved from http://www.wiwiss.fu-berlin.de/en/fachbereich/bwl/pwo/bizer/research/publications/Isele-LinkGeneration-ISSLOD2011.pdf

Isele, R. (2013). Learning Expressive Linkage Rules for Entity Matching using Genetic Programming (Doctoral dissertation).

Schaible, J., & Mayr, P. (2012). Discovering links for metadata enrichment on computer science papers. arXiv preprint arXiv:1212.3677.

Singh, R. (2011). Graphical user interface for silk-a link discovery framework for the web of data. TUDelft.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Ha, Seung Rok, Yim, Jin Hee, & Rieh, Hae-young (2017). A Study on the Procedure for Constructing Linked Open Data of Records Information by Using Open Source Tool. Journal of the Korean Society for Information Management, 34(1), 341-371.

Kim, Yong-Kyeom (2014). Main Application Case and Development Directions of Semantic Web. Journal of East and Central Asian Studies, 25(3), 65-86.

Lee, Kyounguk (2015). LOD InterLinking. Retrieved Retrieved October 19, 2017, from https://www.slideshare.net/ssuser6e1ce5/interlinking-lod

Lee, Sungsook, Park, Ziyoung, & Lee, Hyewon (2017). Expanding the Scope of Identifying and Linking of Personal Information in Linked Data: Focusing on the Linked Data of National Library of Korea. Journal of the Korean Society for Information Management, 34(3), 7-21.

Lee, Yu-Been & Rieh, Hae-Young (2017). A Suggestion of Interface for Ontology-Based Record

Retrieval System. Journal of Korean Society of Archives and Records Management, 17(1), 217-244.

National Information Society Agency (2015). Easy Guide about Linked Open Data. Seoul: National Information Society Agency.

Park, Ok Nam (2012). The Design and Development of Linked Data from Authority Data in National Archives of Korea. Journal of the Korean Biblia Society for Library and Information Science, 23(2), 5-25.

Park, Zi-young (2016). Analyzing the Next-generation Archival Description Standard: "Record in Context" of ICA EGAD. Journal of Korean Society of Archives and Records Management, 16(1), 223-245.

Park, Zi-young (2017). Transition of Archival Description from ISAD(G) to Record in Context Conceptual Model. Journal of Korean Society of Archives and Records Management, 17(1), 93-115.

Yoon, So-Young (2013). A Study on National Linking System Implementation based on Linked Data for Public Data. Journal of the Korean Society for Information Management, 30(1), 259-284.