한국기록관리학회지
21(3), 61-78, 2021.8
http://dx.doi.org/10.14404/JKSARM.2021.21.3.061
pISSN 1598-1487 eISSN 2671-7247

JKSARM
Journal of Korean Society of Archives and Records Management

일본군 '위안부' 지식그래프: 파편화된 디지털 기록의 연결

A Knowledge Graph on Japanese "Comfort Women": Interlinking Fragmented Digital Archival Resources

박하람(Haram Park)[1], 김학래(Haklae Kim)[2]

E-mail: harampark@kakao.com, haklaekim@cau.ac.kr

[1] 제 1저자 중앙대학교 일반대학원 문헌정보학과 문헌정보학전공 석사과정
[2] 교신저자 중앙대학교 사회과학대학 문헌정보학과 교수

논문접수 2021-07-20
최초심사 2021-07-26
게재확정 2021-08-03

ORCID
Haram Park
https://orcid.org/0000-0002-2091-0613
Haklae Kim
https://orcid.org/0000-0002-2616-421X

초 록

일본군 '위안부'에 대한 기록은 민간 기관에서 개별적으로 관리하고 있다. 일부 기록은 디지털 아카이브로 구축되어 온라인으로 접근할 수 있다. 그러나, 디지털 아카이브의 기록은 기관에 따라 메타데이터의 구성과 표현 방식이 다르다. 한편, 기록 사이의 관계를 정의할 수 있는 체계가 미흡하기 때문에, 현재 구축된 일본군 '위안부' 기록은 서로 연결되지 않고 파편적인 형식으로 남아있다. 본 연구는 일본군 '위안부' 디지털 기록을 연계하기 위한 지식 모델을 제안하고, 분산화된 디지털 아카이브의 기록을 통합하여 일본군 '위안부' 지식그래프를 구축한다. 일본군 '위안부' 디지털 아카이브의 메타데이터를 분석하여 공통 요소를 도출하고, 표준 어휘를 적용하여 디지털 기록의 다양한 개체와 개체 사이의 관계를 의미적으로 표현한다. 특히, 흩어져 있는 기록을 연계하고 검색하기 위해 수집한 데이터의 정제가 이루어지고, 외부 데이터를 활용하여 기록의 맥락 정보를 강화하고 있다. 구축된 지식그래프의 검증은 분산된 기록의 탐색 여부를 측정하는 질의를 통해 수행된다. 검증 결과, 지식그래프는 흩어져 있는 기록을 연계하여 검색할 수 있고, 외부데이터로부터의 강화로 기록의 맥락 정보를 풍부하게 제공하며, 의미 기반의 검색을 통해 사용자의 의도에 맞춘 정확한 검색이 가능하다.

ABSTRACT

Records on Japanese "Comfort Women" have been individually managed by private sectors or institutions, and some are provided as digital archives on the Internet. However, records of digital archives differ in the composition and representation of metadata by individual institutions. Meanwhile, there is a lack of a consistent structure to describe the relationships between and among these records, leading to their fragmentation and disconnectedness. This paper proposes a knowledge model for interlinking the digital archival resources and builds a knowledge graph by integrating the records from distributed digital archives. It derives common elements by analyzing metadata from the diverse digital archives and expresses them in standard vocabularies to semantically describe multiple entities and relationships of the digital archival resources. In particular, the study includes the refinement of collected data to search and thread dispersed records and the enrichment of external data to provide significant contextual information of records. An evaluation of the knowledge graph is performed via a query measuring the (dis)connectivity between the distributed records. As a result, the knowledge graph is capable of interlinking and retrieving fragmented records, providing substantial contextual information on the records with external data enrichment, and searching accurately to match the user's intentions through semantic-based queries.

Keywords: 일본군 '위안부' 기록, 지식그래프, 디지털 아카이브, 링크드 데이터
Japanese "Comfort Women" Archives, Knowledge Graph, Schema.org, Digital Archive, Linked Data

▪이 논문은 2021년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.

1. 서 론

일본군 ‘위안부’의 공식 명칭은 일본군 성노예제(Military Sexual Slavery by Japan)로, 제2차 세계대전 동안 일본군이 조직적으로 군위안소를 설치해 점령지와 식민지 여성들을 성노예로 만든 범죄이다(정의기억연대, 2018). 1988년 ‘여성과관광문화 세미나’에서 윤정옥은 최초로 일본군 ‘위안부’ 문제를 소개하고, 1990년 ‘한국정신대문제대책협의회’가 결성되어 일본 정부에 공식적으로 문제를 제기했다(정의기억연대, 2018). 한편, 1991년 8월 고(故) 김학순의 일본군 ‘위안부’ 피해 공개증언을 시작으로, 미국과 일본 등지에서 정신대 관련 문제가 발굴・공개 되어 왔다(이나영, 2010, 42).

일본군 ‘위안부’ 기록은 일본군에 의해 성노예 피해를 당한 사실을 증거한다는 점에서 큰 가치를 지닌다(서연수 외, 2016). 일본 정부가 ‘위안부’ 문제에 조직적으로 관여했다는 사실을 부인하는 상황에서 일본군 ‘위안부’ 기록은 피해를 증명하는 객관적 증거가 된다. 한편, 고령의 피해자들이 사망하고 있으므로 미래 세대의 일본군 ‘위안부’ 문제 해결을 위한 핵심적 증거가 될 수 있다. 따라서 일본군 ‘위안부’ 기록을 체계적으로 발굴・보존하는 기록 관리의 본래 목적과 더불어 기록을 공개하고 공유할 수 있는 체계에 대한 검토도 필요하다.

디지털 아카이브는 지속적으로 보존할 가치를 가진 디지털 자원을 저장하는 물리적 저장소이고, 동시에 가상의 공간에서 디지털 자원을 탐색하고 열람할 수 있는 서비스이다(김학래, 2021). 국내에서 일본군 ‘위안부’와 관련된 디지털 아카이브는 수요시위 아카이브, 아카이브814, 국가기록원에 의해 운영되고 있으며, 개별 아카이브에서 보존하고 있는 기록의 유형, 범위, 제공 형식은 다양하다. 그러나 디지털 아카이브에 접근하고, 기록을 활용하는데 여러 가지 제한이 존재한다. 현재 대부분의 기록은 국내외 포털에서 검색을 지원하지 않고, 해당 아카이브 사이트에 직접 방문해 자체적으로 제공하는 검색 기능을 통해 확인해야 한다. 한편, 현존하는 디지털 기록을 수집 또는 통합할 수 있는 법제도와 라이선스 가이드라인이 미흡하기 때문에, 발굴된 기록은 개별적인 아카이브에 파편적으로 존재하고 있다. 일본군 ‘위안부’ 기록에 대한 참조 모델이나 국가 차원의 양질의 데이터가 구축되지 않은 상황에서, 민간에 존재하는 데이터의 연계는 체계적으로 관리되기 어려운 것이 현실이다.

기술적 측면에서 보면, 분산화된 디지털 아카이브의 기록은 서로 연계되거나 통합되지 않고 있고, 대부분의 디지털 아카이브에서 일본군 ‘위안부’에 대한 기록은 서로 다른 메타데이터 요소로 정의되어 있고, 메타데이터 값의 범위와 유형는 통일되어 있지 않다. 지식 그래프는 일종의 데이터베이스로 인류의 사고 체계를 지식이라는 개념으로 구조적으로 정의한 것이다(김학래, 2017). 지식 그래프에서 지식은 개념에 대한 정의, 개념 사이의 관계로 표현되는데 이를 지식 모델(knowledge model)이라고 한다(Nickel et al., 2015). 수학에서 그래프는 꼭짓점의 집합(set of vertex)과 그 사이를 잇는 변의 집합(a set of edge)으로 구성된다. 즉, 지식 그래프는 개념들의 관계를 그래프 구조의 지식 모델로 표현한 것이다. 지식 그래프는 웹 패러다임의 발전과 함께 성장하였기 때문에, 웹 환경의 다양한 구성요소를 활용하여 구현할 수 있다. 구글, 마이크로소프트, 아마존, 아이비엠 등 주요 ICT 기업들은 자사의 데이터를 대규모 지식 그래프로 구축하고 상용 서비스에 적용하고 있다(Fensel et al., 2020; Zou, 2020). 위키데이터(Wikidata)는 모든 다국어 위키백과의 메타데이터를 하나로 연계된 정보로 제공하는 지식그래프의 사례이다(Vrandečić, 2012). 일본군 ‘위안부’ 기록은 다양한 개체와 개념을 포함하고 있고, 분산적으로 파편화된 상태로 존재하기 때문에, 지식그래프 기술을 적용하여 데이터의 연계와 검색 문제를 해결할 수 있다.

본 연구는 분산적으로 존재하는 일본군 ‘위안부’ 기록을 의미적으로 표현하고 연계하기 위한 방법을 제안한다. 논문의 구성은 다음과 같다. 2장은 기록 관점의 일본군 ‘위안부’에 대해 소개하고, 지식그래프와 관련된 개념과 연구동향을 기술한다. 3장은 국내에서 운영 중인 일본군 ‘위안부’ 디지털 아카이브 사례를 소개한다. 4장은 지식그래프 구축을 위한 과정을 기술한다. 특히 디지털 아카이브에 적용한 메타데이터를 분석하고, 공통 요소를 도출함으로써 분산적으로 존재하는 기록을 연계하기 위한 지식 모델, 데이터 강화에 대해 소개한다. 5장은 구축된 지식그래프에서 의미 질의를 수행하고 기존 디지털 아카이브와 차이점을 설명하며, 6장에서 연구 결과를 정리한다.

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

2. 선행 연구

2.1 기록 관점의 일본군 '위안부'

일본군 '위안부' 문제는 1991년 고(故) 김학순의 '위안부' 피해 기자회견을 시발점으로, 일본군 '위안부' 관련 자료 발굴이 시작되었다(서현주, 2016). 일본에서는 민간이 주도하여 '위안소'제도와 일본군 '위안부' 관련 공문서를 발굴해왔고, 한국, 중국, 대만은 피해자들의 증언과 '위안부' 강제연행 자료, '위안소' 설치 기록 등을 발굴한 성과를 이루었다(김정현, 2020).

그러나 그동안 발굴된 일본군 '위안부' 기록은 여러 대학 연구소와 민간단체, 기관들에 산재되어 존재하여 체계적인 보존·관리가 어려운 실정이다. 산재된 일본군 '위안부' 기록을 통합적으로 관리하려는 시도는 지속적으로 있었다(권미현, 2007; 남영주, 2017; 서연수 외, 2016; 봉지현, 남영준, 2019; 서울대 인권센터 정진성 연구팀, 2018a; 2018b). 국가기록원은 '일본군 위안부 관련 자료'를 국가지정기록물로 선정하여 기록을 통합·관리하고자 했다. 국가지정기록물은 민간기록물 중에서 국가가 영구적으로 보존할 필요가 있는 기록을 지정하여 관리하는 제도로, 국가적 보존가치가 큰 기록의 공적인 관리를 강화하기 위해 마련되었다(국가기록원, 2014a). 국가기록원은 2013년 국가지정기록물 제8호로 3,060점을 선정하였고, 2014년 제8-1호 940점, 제8-2호 125점으로 확장하였다. 국가지정기록물은 생존자 구술기록, 증언영상, 간병일지, 기자회견 영상, 생존자 미술치료 그림 등이 포함된다(국가기록원, 2014b). 그러나 민간이 자율적으로 관리하고 있어 기록의 영구적 보존이 불투명한 상황이다(한국기록전문가협회, 2020).

한편, 각지에 흩어져 있는 일본군 '위안부' 기록을 세계기록유산으로 보존하고자 한 노력도 있다. 2016년, 한국 중심의 8개국 15개 단체가 유네스코 세계기록유산(World's documentary heritage)으로 일본군 '위안부' 기록물을 등재하고자 했다. 유네스코 세계의 기억(UNESCO Memory of the World) 프로그램은 세계사적 가치를 지닌 기록물을 제정하여 기록의 보존을 지원한다(UNESCO, 2021). 등재 신청한 기록물은 총 2,744점으로, 일본군 '위안부' 제도를 증명하거나 '위안부' 생존자들이 생산한 자료 등을 포함한다(신혜수, 2021). 그러나 현재 등재신청은 일본의 방해공작으로 무기한 지연된 상황이다(Shin, 2021).

지속적인 통합 관리와 보존 시도에도 불구하고, 대부분의 일본군 '위안부' 자료는 개별 관리 기관의 자체 기준에 따라 보관되어 있다. 현재 민간기관에서 제공하는 기록은 대부분 디지털화되지 않았기 때문에, 이용자의 접근성과 기록의 탐색성이 낮다. 다수의 관련 기관은 소장기록을 디지털 형태로 전환하여 디지털 아카이브를 구축하고자 한다(남영주, 2017; 윤지현, 2020). 서로 다른 메타데이터를 통해 기술된 일본군 '위안부' 기록은 상호운용성을 저해하고, 기록의 통합 관리를 어렵게 한다. 서연수 외(2016)는 파편화된 일본군 '위안부' 기록의 통합 관리를 위해 통합 메타데이터 스키마를 제안하고 있다. 봉지현과 남영준(2019)은 구술자료의 통합 관리를 위한 메타데이터 요소를 제안하고 있다. 기록의 체계적인 관리를 위해 일관된 메타데이터를 정의하는 것이 필요하지만, 기관마다 관리 체계가 다른 현실을 고려하면 메타데이터의 스키마는 데이터의 연계를 고려하는 수준으로 확장되어야 한다. 본 연구는 공통 메타데이터를 중심으로 일본군 '위안부' 기록을 연계할 수 있는 방안을 제시한다. 더불어 구축된 디지털 아카이브의 접근성과 기록의 탐색성을 강화할 수 있는 방안을 제안한다.

2.2 온톨로지 어휘

온톨로지(ontology)는 "공유된 개념(conceptualization)의 형식적(formal), 명시화된(explicit) 명세(specification)"(Gruber, 1994; Borst, 1997)로 정의된다. Studer와 Benjamins, Fensel(1998)에 따르면, '개념화(conceptualization)'는 실제 세계의 추상적 모델을 의미로 집단이 '공유(share)'하고 있는 지식이다. '형식적(formal)'은 자연어가 아닌 기계로 읽고 처리할 수 있는 언어로 표현된다는 것이며, '명시화(explicit)'는 사용된 개념의 유형과 그 사용에 대한

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

제약이 명확히 정의됨을 의미한다. 온톨로지는 어휘의 내용을 명확히 정의하고 어휘들 간의 관계를 통해 새로운 사실을 추론할 수 있지만, 현재 추세는 범용성을 높여 데이터의 공유와 연결에 중점을 두고 있다(김학래, 2017).

더블린코어 메타데이터(Dublin Core Metadata), Schema.org, FOAF(Friend of a Friend), SKOS(Simple Knowledge Organization System)는 자원을 기술하는 대표적인 어휘다. Schema.org는 웹 상의 구조화된 데이터(structured data)를 웹 단위로 교환 가능(Web-scale exchange)하도록 고안한 일련의 어휘 모음이다(Guha, Brickly, & Macbeth, 2016). 2011년에 구글, 마이크로소프트, 야후와 얀덱스 사에서 공동 개발한 어휘로, 사람, 장소, 사건 등을 포함한 광범위한 주제에 대한 단일한 스키마를 제공하는 것을 목표로 한다(Guha, Brickly, & Macbeth, 2016). 현재 schema.org는 792개의 유형과 1447개의 속성을 갖고 있고 확장기법 (Extension)을 통해 특정 도메인의 어휘를 확장시킬 수 있다(Schema.org, 2021). 도서관과 디지털 아카이브 분야에서도 Schema.org 적용이 확대되고 있다(Jett et al., 2017; Freire, Charles, & Isaac, 2018; Freire et al., 2020). 예를 들어, 서지 표현을 위한 SchemaBibExtend를 제안하여 서지 관련 어휘가 확장되었고(W3C Schema Bib Extend Community Group, 2015), 기록 관련 어휘는 2017년 Schema Architype Extension이 제안되었다(W3C Schema Architypes Community Group, 2015). ISAD(G), ISAAR-CRF, DACS와의 매핑을 고려해 디지털 아카이브를 기술할 수 있도록 확장되었으나(Matienzo, Roke, & Carlson, 2017), 다차원적인 기록 체계를 기술하기에는 한계가 존재한다. 그러나 디지털 아카이브를 링크드 데이터로 쉽게 발행할 수 있고(Matienzo, Roke, & Carlson, 2017), 일반적인 검색 엔진에서 디지털 컬렉션(Collection)의 탐색성을 강화할 수 있다는 강점을 지닌다(Lampron, Mixter, & Han, 2016; Han et al., 2015).

RiC-O(Records in Contexts–Ontology)는 ICA EGAD에서 2017년 개발한 기록 표준 온톨로지 어휘이다. RiC-CM (Records in Contexts–Conceptual Model)은 기록 기술(description)에 필요한 개체(entities)와 상호관계(interrelations)를 표현하는 개념 모델이다(ICA EGAD, 2016). RiC-CM을 기계가 읽고 처리할 수 있는 표현으로 제공한 것이 RiC-O이다(ICA EGAD, 2019). RiC-CM은 기존에 존재하는 기록 표준 ISAD(G), ISAAR(CPF), ISDF, ISDIAH를 통합하고 있다(ICA EGAD, 2016). 따라서 RiC-O는 기록의 다차원적인 체계를 반영하여 기록을 세밀하게 표현하는 것이 가능하다. 대표적으로 프랑스의 ANF(Archives Nationales France, 2021)는 RiC-O를 활용해 여러 기관의 소장 기록을 상호 연결하는 프로젝트를 진행했고, 스위스의 Memobase 포털(Memobase, 2001)은 스위스 기관 80개의 시청각 아카이브 메타데이터를 RiC-O, PREMIS, Ebucore로 표현하여 링크드 데이터로 구축했다(ICA EGAD, 2021). 미국의 SNAC(Social Networks and Archival Context Cooperative, 2021)은 소장기록의 핵심 개체를 RiC-O로 표현하고, SPARQL 엔드포인트(Endpoint)로 기록을 검색할 수 있도록 계획하고 있다(ICA EGAD, 2021).

국내는 RiC 중심의 기록 기술 연구가 활발하다. RiC의 개념과 특성을 분석하여 국내 기록물의 RiC-CM 적용가능성을 논의하는 연구가 진행되었다(박지영, 2017; 박선희, 2019; 신미라, 김익한, 2019; 김수현, 이성숙, 2020; 전예지, 이혜원, 2020). 또한, RiC-O를 활용해 실제 기록물에 적용한 연구도 있다. 정희명과 이성숙(2021)은 국가기록원 소장 기록물 88건을 RiC-O로 표현한 사례를 제공하여 RiC-O가 기록의 맥락을 상세하게 기술할 수 있음을 제시한다. 이유경과 김학래(2020)는 '1997 외환위기아카이브'의 디지털 소장기록 5,381건을 RiC-O 기반의 지식그래프로 구축하고, RiC-O가 기록을 둘러싼 맥락과 개체 관계에 대해 다차원적으로 표현할 수 있다고 제시한다. 한편, RiC-O는 기록의 다차원적인 맥락 기술에 강점을 지니지만, 어휘의 복잡성이 높기 때문에 링크드 데이터 발행을 어렵게 만드는 단점이 있다. 또한, RiC-O는 기존 어휘의 재사용을 고려하지 않고 있어, 어휘 수준의 연계와 상호운용에 제한이 있다. 본 연구는 파편화된 일본군 '위안부' 디지털 기록의 메타데이터를 수집・분석하고, 공통 어휘를 선정하여 이를 위한 지식 모델을 제안한다.

3. 일본군 '위안부' 디지털 아카이브

일본군 '위안부' 기록은 대부분 민간 기관에서 관리하고, 일관성 있는 관리체계 없이 파편화되어 존재한다. 일부

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

기관은 기록의 접근성을 강화하기 위해 디지털 아카이브를 운영하고 있다. 대표적인 일본군 '위안부' 디지털 아카이브는 수요시위 아카이브, 아카이브814, 서울기록원의 일본군 '위안부' 컬렉션(이하 서울기록원), 성평등 아카이브의 일본군 '위안부' 컬렉션(이하 성평등 아카이브), 국가기록원의 국가지정기록물 제8호(이하 국가기록원)가 있다. 각 기관에서 소장하고 있는 기록물의 현황은 <표 1>과 같다.

〈표 1〉 기관별 소장기록 현황

목록	소장 기록물 수(건)	제공 메타데이터 수(개)
수요시위 아카이브	1,085	17
아카이브814	596	20
서울기록원	137	25
성평등 아카이브	408	88
국가기록원	27	20

수요시위 아카이브는 정의기억연대 소관 전쟁과 여성 인권 박물관에서 운영하는 디지털 아카이브이다(전쟁과 여성 인권 박물관, 2020). 1992년 1월부터 시작된 "일본군성노예제 문제해결을 위한 정기 수요시위"에 대한 기록을 담고 있다(전쟁과 여성 인권 박물관, 2020). 수요시위 아카이브는 총 1,085건의 기록을 보유하고 있고, 기록에 대한 메타데이터는 자체적으로 정의한 17개 항목으로 구성된다. 모든 기록은 JPEG, PDF, MP4 형식으로 원문이 공개되며, 키워드 기반의 통합검색과 상세검색 서비스를 제공한다.

아카이브814는 여성가족부 산하 일본군 '위안부' 문제연구소에서 운영하는 아카이브이다. 고(故) 김학순이 일본군 '위안부' 피해 사실을 처음 공개적으로 증언한 날인 8월 14일(일본군 '위안부' 기림의 날)을 기념해 아카이브가 개설되었다(일본군 '위안부' 문제연구소, 2020). 아카이브814는 국내외 법정기록, 일본군 '위안부' 관련 공문서, 주제별 컬렉션, 연표, 소장 도서 목록 등 총 596건의 기록을 20개의 메타데이터 항목으로 표현하고 있다. 기록은 PDF 형태로 원문이 제공된다. 기능적 측면에서, 아카이브814는 키워드 기반의 통합검색 서비스를 제공하고, 서울기록원에서 소장하고 있는 일본군 '위안부' 컬렉션 기록을 외부 링크로 연결하여 검색할 수 있다.

서울기록원은 서울대 정진성 연구팀이 서울시의 지원을 받아 발굴・수집한 일본군 '위안부' 기록을 컬렉션 형태로 제공한다(서울기록원, 2019). 서울대 정진성 연구팀은 제2차 세계대전 당시 연합군이 생산한 자료 중 일본군 '위안부'와 위안소 존재를 증명하는 자료를 발굴・수집하여 서울기록원에 이관하였다. 총 137건의 기록이 제공되며, 자체적으로 정의한 25개 메타데이터로 기술되어 있다. 원문은 PDF, JPG, MP4의 형태로 공개된다. 서울기록원은 '해당 기록 내 검색'을 통해 일본군 '위안부' 기록을 검색할 수 있는 통합검색 기능을 제공한다.

성평등 아카이브는 서울시여성가족재단의 성평등 도서관 '여기' 관할 아카이브로, 일본군 '위안부' 관련 자료를 컬렉션 형태로 제공한다(성평등 아카이브, 2019). '정신대 문제'와 민간 차원의 국제 인권 법정이었던 '2000년 일본군 성노예 전범 여성 국제 법정'을 중심으로 기록을 제공한다. 기록은 총 408건이 제공되고, 88개의 메타데이터로 기술된다. 모든 기록은 PDF 형태로 제공된다. 특히, 성평등 아카이브는 기록물에 대한 메타데이터를 '목록정보받기'를 통해 CSV 형태로 제공한다. 성평등 아카이브가 소장하고 있는 모든 기록에 대한 '일반 검색' 서비스는 제공하지만, 일본군 '위안부' 관련 자료에 국한된 검색 서비스는 제공하지 않는다.

국가기록원은 일본군 '위안부' 관련 기록물을 국가지정기록물 제8호로 지정하고 있다(국가기록원, 2013). 나눔의 집과 정신대할머니와 함께하는 시민모임에서 소장하는 기록물 대략 3,060점이 국가지정기록물로 선정되었다. 국가기록원은 주요기록물로 27점을 선정하여 디지털화된 기록물과 20개의 메타데이터 설명 정보를 제공한다. 기록은 공개에 한하여 국가기록원 뷰어(Viewer) 형태로 원문을 제공한다. 키워드 기반의 검색 서비스를 제공하며 기록물 형태, 기록물명, 생산기관, 생산연도에 따라 검색이 가능하다.

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

본 연구는 파편화된 일본군 ‘위안부’ 디지털 기록을 연계하기위해 지식그래프 기술을 활용한다. 소장기록은 공통 메타데이터로 구축하고, 기록에 기술된 자원은 Schema.org를 중심의 기계가 읽고 처리할 수 있는 형식으로 표현한다. 기록, 관련 인물과 조직은 URI를 부여하여 개체로 식별하고, 의미 수준에서 상호운용성을 확보하여 서로 다른 디지털 아카이브의 기록과 연계한다.

4. 지식그래프 설계와 구축

4.1 데이터 수집

본 연구는 일본군 ‘위안부’ 지식그래프를 구축하기 위해 디지털 아카이브의 소장기록과 메타데이터 항목을 수집한다. 수집된 기록물은 총 2,253건이다(수요시위 아카이브 1,085건, 아카이브814 596건, 성평등 아카이브 408건, 서울기록원 137건, 국가기록원 27건). 아카이브814의 3건, 국가기록원 2건은 기록 정보가 충분하지 않아 제외하고, 기록물 2,248건을 지식그래프로 구축한다.

개별 아카이브에서 정의한 메타데이터 항목은 서로 다르다. 일본군 ‘위안부’ 지식그래프 구축을 위해 메타데이터 항목의 특성을 검토하고 활용 대상이 되는 항목을 선정해야 한다. 메타데이터 선정은 다음 3가지 기준을 따른다. 첫째, 공통으로 적용된 메타데이터를 추출하고 활용한다. 개별 아카이브에서 2개 이상 공통으로 사용된 메타데이터는 “제목”, “설명”, “식별자(등록번호 또는 관리번호)”, “생산일자”, “생산자”, “라이선스”, “언어”, “기록유형” 등이 있다. 둘째, 메타데이터 항목은 있으나 실제 데이터가 없는 메타데이터는 구축 범위에서 제외한다. 예를 들어, 성평등 아카이브는 88개의 메타데이터 항목을 제공하지만, 60개의 항목은 데이터 값이 존재하지 않는다. 이런 경우, 메타데이터 선정에서 제외한다. 셋째, 서로 다른 아카이브 간의 ‘위안부’ 기록물의 연계를 표현하는데 필요한 메타데이터 항목은 추가로 정의한다. 예를 들어, 기록물은 각기 다른 아카이브 관리 기관에서 관리하며, 이를 기술하기 위한 “아카이브 관리자” 항목이 필요하다. 메타데이터 항목에 관련 정보가 없을 경우, “아카이브 관리자” 항목을 추가하여 기록물 관리기관 정보를 기술할 수 있게 한다. <표 2>는 지식그래프 구축에 활용할 메타데이터 항목으로 개별 아카이브에서 공통 요소를 선정한 결과이다.

〈표 2〉 소장 기록의 메타데이터 목록

목록	활용 메타데이터 목록	활용 메타데이터 수(개)
수요시위 아카이브	제목, 설명, URL*[1], 형태분류, 첨부파일*, 시기분류, 생산일자, 생산자, 출처분류, 태그, 주제분류, 관련 용어, 라이선스*, 원문보기*, 아카이브 관리자*, 식별자*	15
아카이브814	제목, 제목(원문), 범위와 내용, 번역문 보기, 등록번호, 생산일자, 생산기관(생산자), 분량, 언어, 활용조건(저작권), 기증자/수집처, 형태, 시기, URL*, 아카이브 관리자*, 원문보기*, 다운로드*	17
서울기록원	제목, 다운로드*, 식별번호, 소장처, 기록유형, 기술, 기여자, 연관정보, 수집/이관 기관, 언어, 이용유형, URL*, 원문보기*	13
국가기록원	건 제목, 생산기관, 관리번호, 생산연도, 문서유형, 관리기관, 기록물형태, 기록물건 등록번호, 페이지정보, 발행일, 서고정보, 서비스권자(아카이브 관리자), 원문보기, URL*, 라이선스*	14
성평등 아카이브[2]	dc:Identifier, dc:Title, dc:Description, itm:기록물유형, itm:기록물형태, itm:생산자, itm:날짜, itm:크기/분량, itm:언어, itm:발행/출판, itm:기증자, itm:관련 인물, itm:관련 사건, itm:관련 단체, itm:원본소장처, file, 아카이브 관리자*	18

1) 추가로 정의한 메타데이터 항목은 *를 붙여 표기한다.
2) dc는 Dublin Core의 약어이고, itm은 Item Type Metadata의 약어이다.

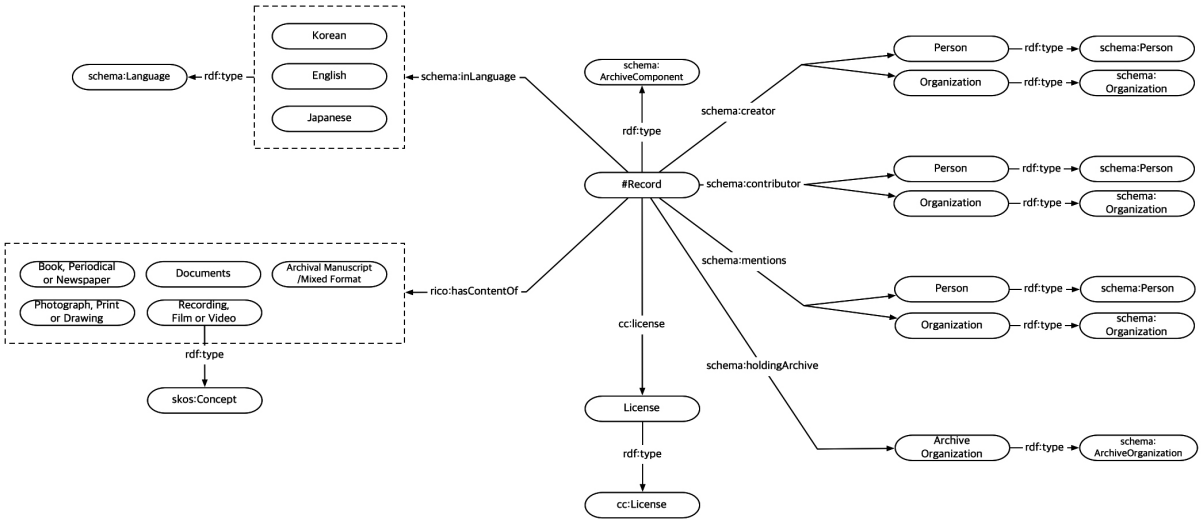
http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

한편, 파편화된 기록물의 의미적 연계를 위한 데이터 요소를 선정해야 한다. 기록의 연계는 오픈리파인(OpenRefine)을 이용해 다음의 정제 과정을 통해 수행된다.

- 인물, 조직, 사건 명칭과 형식을 통일한다. 예를 들어, "한국정신대문제대책협의회"는 한글 명칭과 함께 "Korean Council for Women Drafted by Military Sexual Slavery by Japan"과 같이 영문 표기가 존재한다. 따라서 동일한 개체는 명칭을 통일하여 같은 개체임을 알 수 있도록 한다. 또한 서울기록원은 단체를 "[조직/단체] 서울대학교 정진성연구팀, 2015~"의 형식으로 기록하고, 아카이브814에서는 인물을 "야마가타현 지사 다케이 군지(山形縣知事 武井群嗣)"의 형식으로 기록한다. 즉, 인물이나 조직 명칭 외에 분류기준, 인물의 직위 등의 부가적인 정보와 함께 표기하고 있으므로 인물과 조직 명칭만 표기하여 형식을 통일한다.
- 공통 메타데이터 항목명과 데이터 값 형식을 통일한다. 기록유형은 모든 아카이브의 기록물에서 제공하는 메타데이터로서, 기록유형의 분류기준의 명칭을 통일하면 기록유형에 따라 연계된 모든 기록물을 분류할 수 있다. 기록유형은 개별 아카이브에서 다른 명칭을 사용하고 있어 "기록유형"으로 메타데이터 이름을 통일한다. 기록유형의 분류기준 명칭 또한 상이하므로 "도서/간행물류", "문서류", "박물류", "사진그림류", "영상음성류"로 명칭을 정의한다. 그 외에 "생산일자", "페이지 정보" 등의 메타데이터 항목명과 데이터 값 형식을 통일하여 모든 기록물에서 형식의 일관성을 확보한다.
- 데이터 값의 오류나 공백 등을 수정한다. 예를 들어 인물 명칭이 "Gabrelle Kirk McDonald"로 잘못 표기된 경우, "Gabrielle Kirk McDonald"로 정정하여 표기한다.

4.2 온톨로지 어휘 매핑

일본군 '위안부' 기록을 기술하기 위한 어휘는 어휘의 재사용을 원칙으로 한다. 일본군 '위안부' 기록은 더블린 코어로 표현하고, Schema.org 어휘를 사용하여 검색 엔진에서 탐색할 수 있도록 설계한다. Schema.org 어휘로 표현하지 못하는 경우, RiC-O와 더블린 코어, SKOS 어휘를 적용한다.



〈그림 1〉 일본군 '위안부' 지식모델 구조

일본군 '위안부' 지식모델의 기본 구조는 <그림 1>과 같다. 기록(#Record)은 schema:ArchiveComponent의 인스턴스로서, 각 아카이브에서 소장하고 있는 기록을 의미한다. 기록을 표현하는 주요한 정보는 크게 2가지로 나누어 표현

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

된다. 첫째, 기록과 관련된 인물과 조직 정보를 표현한다. 기록을 만든 생산자(schema:creator), 기록이 생산되는데 도움을 준 기여자(schema:contributor), 기록과 관련된 정보(schema:mentions), 기록을 소장하고 있는 아카이브 관리자(schema:holdingArchive) 등이 있다. 생산자, 기여자, 관련 정보는 인물(schema:Person)과 조직(schema:Organization)으로 나누어 클래스를 부여한다. 아카이브 관리자는 schema:ArchiveOrganization 클래스로 표현된다. 둘째, 기록의 특성과 관련된 정보를 표현한다. 기록의 작성 언어(schema:inLanguage), 기록의 유형(rico:hasContentOf), 이용조건이 표기된 라이선스(cc:license) 등을 표현한다. 언어는 한국어, 일본어, 영어를 포함하고 schema:Language 속성으로 표현한다. 기록유형은 '도서/간행물류', '문서류', '박물류', '사진그림류', '음성영상류'로 나누어 skos:Concept 클래스로 매핑한다. 한편, 라이선스는 cc:License 클래스를 부여하나, 라이선스 유형에 따른 조건까지 함께 표현한다. 예를 들어 공공누리 제4유형은 출처 표기(cc:Notice)를 해야 하고(cc:require), 상업적 이용(cc:CommercialUse)이 불가능하며(cc:prohibits), 2차 저작물의 작성(cc:DerivativeWorks)이 불가능하다(cc:prohibits)로 표현된다.

지식모델 구조에 따라 공통 메타데이터 항목과 어휘를 매핑한 현황은 <표 3>과 같다. 개별 아카이브의 메타데이터 특성을 분석하여 공통 요소별로 매칭하고, 공통 요소에 지식모델에서 정의한 클래스와 속성 정보를 부여한다. 예를 들어, '1992년 제1차 정신대문제 아시아연대회의 취지문' 기록(schema:ArchiveComponent)은 일본군 '위안부' 문제연구소(schema:ArchiveOrganization)에 의해 관리되며(schema:holdingArchive), 한국정신대문제대책협의회(schema:Organization)에 의해 생산되었다(schema:creator). 기록유형은 문서류(skos:Concept)이며(rico:hasContentOfType), 한국어(schema:Language)로 표기되었다(schema:inLanguage). 해당 기록은 공공누리 제4유형(cc:License)의 라이선스로 표기된다(cc:license).

4.3 데이터 강화(Data Enrichment)

수집된 일본군 '위안부' 데이터는 외부 데이터와 연계하여 추가적인 정보를 표현할 수 있다. 본 연구에서 인물, 조직 정보는 위키데이터에 있는 메타데이터와 연계하여 데이터를 확장하고 개체 사이의 관계를 강화한다. 데이터 강화는 오픈리파인의 RDF extension 3.4.1 기능을 활용하고, 추가된 모든 정보는 RDF 변환을 통해 지식그래프에 포함된다.

데이터 강화는 데이터의 선정, 일치하는 개체의 식별, 추출, 변환 등 일련의 과정을 통해 수행된다.

- 일본군 '위안부' 기록에 기술된 인물, 조직 용어를 담은 사전을 구축한다. 이를 위해 메타데이터 항목에 기술된 인물, 조직 용어를 수집한다. 사전 구축에 활용된 메타데이터 항목과 수집된 개체 수는 <표 4>와 같다. 1차적으로 수집된 개체 수는 총 654개이다. 개체 사전에서 중복 개체를 제거하고 인물, 조직으로 개체를 분류한다. 원본 데이터에서 조직, 인물 정보를 분류하여 제공하지 않기 때문에, 분류는 수작업이 포함된다. 동일한 개체가 여러 개의 언어로 표현되었을 경우, 개체의 한국어 이름을 우선순위로 두고 다른 언어의 이름은 부제로 기록한다. 결과적으로 구축된 사전은 인물 160개, 조직 302개의 용어를 포함한다.
- 데이터 조화(reconciliate) 기능으로 위키데이터에서 사전 데이터와 일치하는 개체를 탐색한다. 1차적으로 자동으로 위키데이터와 일치하는 개체를 탐색한다. 그러나 자동으로 위키데이터와 연결할 경우 동명이인과 연결되거나 위키데이터와 일치하는 개체가 존재함에도 연결되지 않을 수 있다. 따라서 2차적으로 연구자가 위키데이터와 일치하는 개체를 매칭시킨다. 최종적으로 인물은 68%(109개), 조직은 29%(88개)의 일치율을 보인다.
- 위키데이터와 일치하는 개체의 데이터를 강화한다. 위키데이터와 일치하는 개체는 위키데이터에 존재하는 속성 정보를 가져올 수 있다. 이는 사전 데이터에 없는 속성 정보를 추가할 수 있어 보다 풍부한 인물, 조직 정보를 제공한다. 예를 들어, "김복동(Kim Bok-dong)"는 위키데이터(Q16175111)와 연계되어 시민권(country of citizen), 직업(occupation), 출생 장소(place of birth), 성별(sex or gender) 등의 정보를 가져올 수 있다. 결과적으로 인물에서는 6개, 조직에서는 3개의 속성 정보를 가져온다.

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

〈표 3〉 주요 메타데이터 항목의 어휘 매핑 현황

공통 요소	수요시위 아카이브	아카이브814	서울기록원	국가기록원	성평등 아카이브	속성(Attribute)	개체(Entity)	값(Value)
제목	제목	제목	제목	건 제목	dc:Title	schema:title	schema:ArchiveComponent	xsd:string
식별자	식별자	등록번호	식별번호	관리번호	dc:Identifier	schema:Identifier	schema:ArchiveComponent	xsd:string
설명	설명	범위와 내용	기술		dc:Description	schema:description	schema:ArchiveComponent	xsd:string
생산일자	생산일자	생산일자		생산연도	itm:날짜	schema:dateCreated	schema:ArchiveComponent	xsd:dateTime
생산자	생산자	생산기관(생산자)		생산기관	itm:생산자	schema:creator	schema:ArchiveComponent	schema:Person; schema:Organization
라이선스	라이선스	활용조건(저작권)	이용유형	라이선스		cc:license	schema:ArchiveComponent	cc:License
아카이브 관리자	아카이브 관리자	아카이브 관리자		서비스권자	아카이브 관리자	schema:holdingArchive	schema:ArchiveComponent	schema: ArchiveOrganization
접근 URL	URL	URL	URL	URL		schema:sameAs	schema:ArchiveComponent	schema:URL
원문보기	원문보기	원문보기	원문보기	원문보기	file	schema:mainEntityOfPage	schema:ArchiveComponent	schema:URL
다운로드	첨부파일	다운로드	다운로드			schema:downloadUrl	schema:ArchiveComponent	schema:URL
기록유형	형태분류	기록유형	기록유형	기록물형태	itm:기록물유형	rico:hasContentOfType	schema:ArchiveComponent	skos:Concept
기록물 형태		기록형태		문서유형	itm:기록물형태	rico: hasDocumentaryFormType	schema:ArchiveComponent	skos:Concept
분량		분량		페이지정보	itm:크기/분량	schema:numberOfPages	schema:ArchiveComponent	xsd:nonNegativeInteger
언어			언어		itm:언어	schema:inLanguage	schema:ArchiveComponent	schema:Language
시기분류	시기분류	시기				schema:temporalCoverage	schema:ArchiveComponent	xsd:string
관련 정보	관련 용어		연관정보		itm:관련 인물; itm:관련 단체; itm:관련 사건	schema:mentions	schema:ArchiveComponent	schema:Person; schema:Organization; schema:Event
기여자		기증자/수집처	기여자, 수집/이관 기관		itm:기증자	schema:contributor	schema:ArchiveComponent	schema:Person; schema:Organization

https://jksarm.koar.kr
http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

〈표 4〉 사전 구축에 활용된 메타데이터 항목과 개체 수

목록	메타데이터 항목	개체 수(개)
수요시위 아카이브	생산자, 관련 용어, 아카이브 관리자	16
아카이브814	생산자, 기증자/수집처, 아카이브 관리자	19
서울기록원	소장처, 수집/이관 기관, 기여자, 연관정보, 아카이브 관리자	14
국가기록원	서비스권자(아카이브 관리자), 부서명, 서고 정보, 생산기관	14
성평등 아카이브	itm:생산자, itm:발행/출판, itm:관련 인물, itm:관련 단체, itm:기증자, itm:기술자, itm:원본소장처(아카이브 관리자)	18

- 데이터 강화를 통해 확장된 정보는 적절한 온톨로지 어휘와 매핑한다. 인물에 대한 6개의 정보와 조직에 대한 3개의 정보를 온톨로지 어휘로 표현한다. 인물(#Person)은 schema:Person의 인스턴스로, 일본군 '위안부'와 관련된 활동이나 업무를 수행한 사람을 의미한다. 인물에 대한 자세한 정보는 시민권(schema:nationality), 출생 장소(schema:birthPlace), 성별(schema:gender), 직업(schema:hasOccupation)으로 표현된다. 한편, 조직(#Organization)은 schema: Organization의 인스턴스로, 일본군 '위안부' 활동이나 업무에 관여한 조직을 의미한다. 조직의 자세한 정보는 상부 조직(schema:parentOrganization), 활동 지역(schema:areaServed), 물리적 위치(schema:location)의 속성으로 표현된다.
- 사전의 모든 개체에 고유 식별자 URI를 부여한 후, RDF로 변환한 사전 데이터를 기록 데이터세트에 조화(reconciliate)시킨다. 일치된 개체는 사전 데이터와 동일한 URI를 가진다. 사전 데이터를 RDF로 변환하게 되면, 위키데이터로부터 강화된 정보를 담은 관계 중심의 그래프로 전환된다. 강화된 데이터세트를 RDF 형식의 그래프로 추출한 후, 기록 데이터세트의 <표 4> 메타데이터 항목에 조화시킨다. 일치된 개체는 위키데이터로부터 강화한 사전 데이터를 연계할 수 있게 된다.

4.4 지식그래프 변환

지식 모델이 적용된 데이터세트는 RDF 그래프 형태로 변환하여 그래프 데이터베이스에 저장한다. 그래프 데이터베이스는 RDF 표준 질의 언어 SPARQL로 질의할 수 있는 GraphDB를 사용한다. 일본군 '위안부' 지식그래프 구축 결과는 <표 5>와 같다. 일본군 '위안부' 기록 데이터세트는 명시적 진술문 46,211개와 추론된 진술문 144개로 표현된다. 소장기록에는 직접 RDF 형태로 구축한 크리에이티브 커먼스(Creative Commons)와 공공누리 라이선스가 포함된다. 라이선스는 명시적 진술문 29개와 추론된 진술문 76개로 표현된다. 위키데이터로부터 강화한 사전 데이터세트는 명시적 진술문 2,144개와 추론된 진술문 102개로 표현된다. 따라서 최종 구축된 결과는 명시적 진술문 47,319개와 추론된 진술문 172개이다.

한편, 데이터 강화를 통해 확장된 개체 결과는 다음과 같다. 인물 정보는 6개의 속성 정보가 강화되어 총 87개의 개체가 추가되었다. 조직 정보는 3개의 속성 정보가 강화되어 총 37개의 개체가 추가되었다. 중복 개체 5개를 제외하고 총 119개의 개체가 위키데이터를 통해 확장되었다.

〈표 5〉 일본군 '위안부' 지식그래프 구축 결과

구분	명시적 진술문(개)	추론된 진술문(개)
소장기록	46,211	144
사전(인물・조직)	2,114	102
라이선스	29	76
합계(중복 제외)	47,319	172

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

5. 의미적 연계의 검증

서로 다른 아카이브에 있는 기록은 의미적으로 연계되어 검색이 가능해진다. <표 6>은 의미 검색을 위해 질의와 질의 목적을 구분하고 있다. Q1, Q2, Q3은 다양한 검색 조건에 따라 파편화된 기록을 연계하여 검색하는 목적이고, 활용도가 높은 검색 조건을 세부 내용으로 선정한다. Q4와 Q5는 기록의 맥락 정보가 충분히 제공되는지 판단하기 위한 질의이며, 일본군 '위안부' 기록의 이해에 중요한 조직과 인물을 선택한다. 모든 질의문은 RDF 표준 질의 언어 SPARQL을 사용한다. 예를 들어, <표 7>은 Q3의 질의문으로, 일본군 '위안부' 기록 중 1990년부터 1994년에 생산된 기록을 오래된 순으로 정렬한다. 이때, 모든 개체의 값은 rdf:type과 정확히 일치하는 대상이며, 물리적인 위치에 관계 없이 URI 기반으로 개체를 식별하고 검색 결과에 포함시킨다.

〈표 6〉 질의문 목록

질의문	설명	카테고리
Q1	일본군 '위안부' 기록을 모두 선택한다.	연계성, 탐색성
Q2	기록유형이 '문서류'인 기록을 모두 선택한다.	연계성, 탐색성
Q3	1990년에서 1994년에 생산된 기록을 가져와 오름차순으로 정렬한다.	연계성, 탐색성
Q4	'여성가족부'에 대한 정보를 모두 선택한다.	상호운용성
Q5	'얀 루프 오헨 (Jan Ruff-O'Herne)'에 대한 정보를 모두 선택한다.	상호운용성

〈표 7〉 Q3 SPARQL 질의문 예시

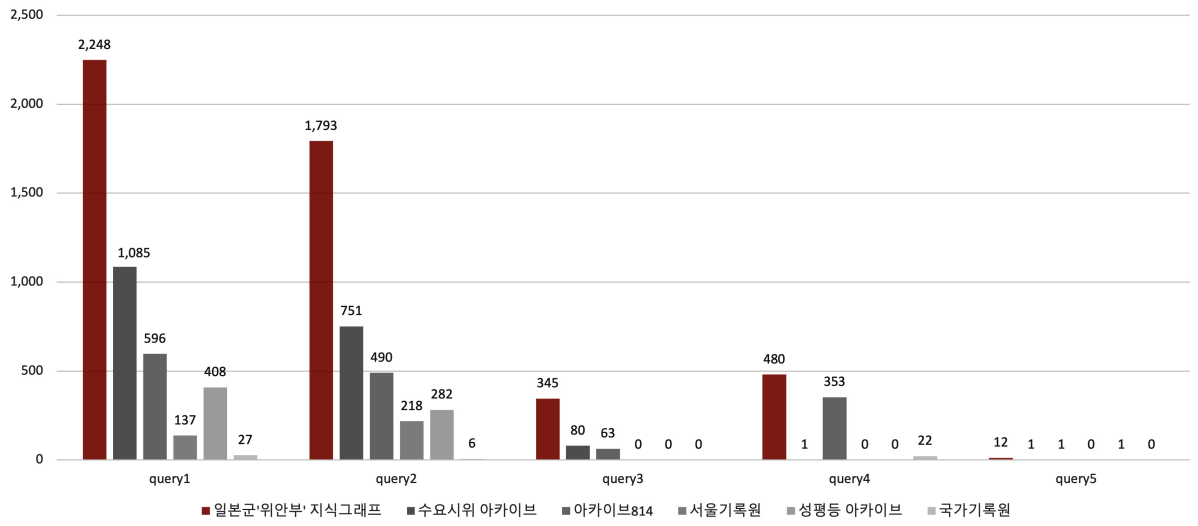
```
PREFIX schema: 〈http://schema.org/〉
PREFIX rdfs: 〈http://www.w3.org/2000/01/rdf-schema#〉
PREFIX xsd: 〈http://www.w3.org/2001/XMLSchema#〉

SELECT ?title ?date ?ArchiveOrganizationName
WHERE {
    ?record rdf:type schema:ArchiveComponent;
            schema:title ?title;
            schema:dateCreated ?date;
            schema:holdingArchive ?ArchiveOrganization .
    ?ArchiveOrganization rdfs:label ?ArchiveOrganizationName
    FILTER (?date 〉= '1990-01-01'^^xsd:date && ?date 〈= '1994-12-31'^^xsd:date)
}
ORDER BY ?date
```

개별 질의를 수행한 결과는 지식그래프의 사용 목적을 명확히 설명하고 있다. 첫째, 일본군 '위안부' 지식그래프는 한 번의 질의로 분산적으로 운영되는 아카이브의 기록을 탐색할 수 있다. 개별 아카이브는 소장하고 있는 기록만 검색할 수 있지만, 지식그래프는 분산된 기록도 함께 검색되므로 기록의 통합 관리가 가능해진다. <그림 2>는 동일한 질의문을 지식그래프와 디지털 아카이브에 검색한 결과이다. Q1-Q3 결과에서 보듯이, 지식그래프는 개별 아카이브의 기록을 모두 포함하고 있다. 지식그래프의 Q1은 구축 범위에 포함된 2,248건이 모두 검색되는 한편, Q2에서는 46건, Q3에서는 202건 더 많은 검색 결과가 도출된다. 이와 같은 결과는 SPARQL로 의미 수준에서 다양한 조건의 질의가 가능해졌기 때문이다. 예를 들어, 성평등 아카이브는 1990-1994년에 생산된 기록을 169건

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

소장하고 있으나, 이에 대한 검색을 지원하지 않는다. 따라서 Q2와 Q3를 통해 기록유형이나 생산일자에 따라 기록을 검색할 수 없다. 반면, 지식그래프는 기록유형을 rico:hasContentTypeOf로, 생산일자는 schema:dateCreated로 의미적 수준에서 정보를 표현하기 때문에 관련된 169건의 기록을 검색할 수 있다.

둘째, 일본군 '위안부' 지식그래프는 의미 기반 검색을 통해 사용자의 의도에 맞춘 정확한 검색이 가능하다. Q4는 '여성가족부'와 관련된 정보를, Q5는 '얀 루프 오헨'과 관련된 정보를 검색한다. 개별 아카이브는 검색 키워드와 일치하는 기록을 검색한다. 따라서, '여성가족부'와 '여성부(2000-2005, 2008-2010년)'는 동일한 개체로 인식하지 못한다. '얀 루프 오헨'도 마찬가지로, '얀 루프 오헤른', '얀 루프 오헨 할머니', 'Jan Ruff-O'Herne', 'Jan Ruff O'Herne'을 모두 다른 개체로 간주한다. 동일한 개체가 상이한 문자열로 표현된 기록은 검색 시스템에 따라 검색 결과와 정확성이 다를 수 있다. 지식그래프는 기록, 관련 인물, 조직을 개체로 정의하고, 문자열이 다른 정보를 동일한 개체로 식별하고 연계하는 과정을 포함한다. 지식그래프에서 '여성가족부'와 '여성부'는 조직의 유형이고, '얀 루프 오헤른'는 사람으로 표현된다. 모든 개체는 문자열의 표현이 아닌 의미적 수준에서 동일한 개체로 인식되어 동일한 URI 체계를 부여한다. 즉, 특정 개체를 다국어로 표현하는 경우도 동일한 개체로 식별할 수 있다. 또한 지식그래프에서 개별 기록은 다양한 개체와 의미적 관계를 갖고 있다. 예를 들어, 여성가족부는 "2017년 여성가족부 일본군 '위안부' 피해자 문제에 대한 보고서" 기록을 생산했고(schema:creator), "1944년 신문기사 일본군 '위안부'들(Comfort Girls)" 기록의 생산에 기여했다(schema:contributor)라고 표현된다. 개체 간의 관계가 의미적으로 연결되므로, 단순한 키워드 검색을 넘어 의미 관계의 탐색이 가능해진다. 그 결과, 지식그래프에서 기존 아카이브보다 Q4에서는 104건, Q5에서는 9건 더 많은 검색 결과를 보인다.

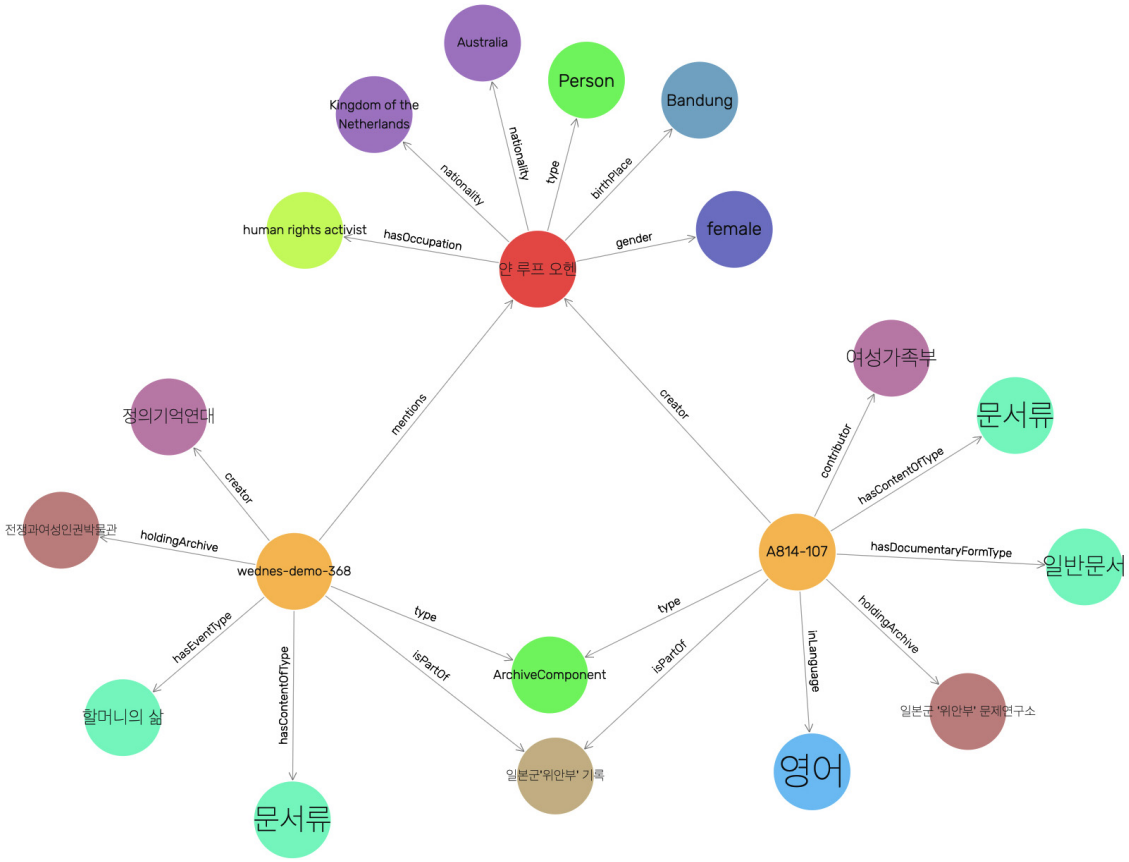


〈그림 2〉 질의문에 따른 검색 결과(건)

셋째, 일본군 '위안부' 지식그래프는 외부 데이터와 연계되어 기록의 맥락 정보를 풍부하게 제공한다. 기존 아카이브는 메타데이터에 기술된 인물, 조직, 사건에 대해 구체적인 정보를 제공하지 않고 있다. 메타데이터에 대한 정보는 기록의 이해에 풍부한 맥락을 제공하기 때문에 중요하다. 예를 들어, '얀 루프 오헨'은 일본군에 의해 성노예 피해를 당한 네덜란드계 호주인이다. 1992년 일본군에 의해 성노예 피해를 당한 사실을 고백한 이후로 인권운동가로서 활발한 평화·인권 운동을 펼쳐왔다. 만약 얀 루프 오헤른이 일본군 '위안부' 피해자였다는 사실을 모른다면, 아카이브814에서 제공하는 "2007년 미 국회 결의한 121를 지지하는 얀 루프-오헤른의 편지" 기록을 충분하게 이해하기는 어렵다. 그러나 <그림 3>과 같이 일본군 '위안부' 지식그래프는 위키데이터로부터의

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

강화로 인물이나 조직에 대한 풍부한 정보를 제공한다. 얀 루프 오헨의 정보는 위키데이터와 의미적으로 연계되어 개체 유형(인물), 성별, 국적, 직업 등의 속성 정보가 강화되었다. 네덜란드계 호주 국적의 여성인 얀 루프 오헨이 인권운동의 활동가였다는 사실은 앞서 제시한 기록을 보다 입체적으로 이해할 수 있도록 돕는다. 또한 지식그래프에서는 얀 루프 오헨이 생산하거나 기여한 다른 기록과 연결된다. 수요시위 아카이브의 기록(wednes-demo-368)은 얀 루프 오헨과 관련 정보(schema:mentions)로 연결된다. 아카이브814의 기록(A814-107)은 얀 루프 오헨을 기록의 생산자(schema:creator)로 연결한다. 따라서 외부데이터로부터 연계된 속성 정보와 기록, 인물, 조직 개체들 간의 연결은 다양한 맥락정보를 통해 기록의 효과적인 이해를 돕는다.



〈그림 3〉 '얀 루프 오헨' 관련 기록 시각화 사례

6. 결 론

본 연구는 일본군 '위안부' 기록의 연계를 위해 지식 모델을 제안하고, 분산적인 디지털 아카이브에 존재하는 기록을 통합하여 지식그래프를 구축했다. 지식 모델은 확장성 있는 표준 어휘 Schema.org, RiC-O, SKOS 등을 중심으로 기술하여 의미적인 수준에서 상호운용성을 확보하였다. 또한 기록과 관련 인물, 조직을 관계 중심으로 재구성하여 보다 입체적으로 기록을 표현하였다. 그 결과, 일본군 '위안부' 지식그래프는 분산되어 있던 일본군 '위안부' 디지털 기록을 연계하여 통합적으로 검색할 수 있고, 기록의 탐색성을 향상시킬 수 있다.

http://dx.doi.org/10.14404/JKSARM.2021.21.3.061

일본군 '위안부' 지식그래프는 디지털 아카이브의 기록을 지식 모델을 기반으로 구축함으로써 다음의 강점을 갖는다. 첫째, 흩어져 있는 기록을 연계하여 검색할 수 있다. 하나의 질의문으로 외부의 기록을 검색할 수 있고, 의미 수준에서 다양한 조건으로 기록을 검색할 수 있다. 특히, 기록이 분산적으로 존재하는 환경에서 연계된 기록에 접근할 수 있는 장점이 있다. 둘째, 기록에 대한 풍부한 맥락 정보를 제공하여 기록의 이해를 돕는다. 의미 기반 어휘는 직접 데이터를 구축하지 않고도 자동적으로 의미가 동일한 개체 정보를 강화할 수 있다. 외부데이터로부터의 강화를 통해 기록에서 제공하지 않은 정보까지 습득할 수 있어 기록의 이해에 효과적이다. 셋째, 의미 기반 검색을 통해 사용자의 의도에 맞춘 정확한 검색이 가능하다. 디지털 기록에 기술된 정보가 의미 중심의 관계로 재구성되어 키워드 기반 검색보다 더욱 정확하게 검색할 수 있다.

한편, 기록의 연계와 활용을 극대화하기 위해서는 고려해야 할 사항이 있다. 첫째, 기록은 정확하고 풍부한 의미 정보를 포함해야 하고, 이에 대한 메타데이터의 선정과 정의가 필요하다. 수집한 디지털 아카이브는 평균 16개의 메타데이터 항목을 제공하고 있지만, 기관에 따라 적용하는 메타데이터 항목이 다르다. 또한, 메타데이터의 값이 부정확하거나 공백으로 존재하는 사례도 다수 발견되었다. 기록에 대한 메타데이터는 기록을 이해할 수 있는 지표이므로, 충실한 메타데이터 기술은 기록의 연계와 공유, 재사용성을 강화할 수 있는 근간이 된다. 둘째, 기록의 활용을 위해 이용조건을 명확히 제공하는 것이 필요하다. 대부분의 일본군 '위안부' 기록은 이용조건에 대한 명확한 라이선스를 제공하고 있지 않다. 명확하지 않은 이용조건은 기록의 활용을 제한한다. 국제 또는 국내의 표준 라이선스를 사용하여 국내외 이용자들이 이용조건을 판단할 수 있도록 하고, 이를 메타데이터 항목으로 제공하는 것이 필요하다. 더 나아가 기계가 읽고 처리할 수 있는 형식으로 라이선스가 제공되면, 기록의 활용성을 더욱 높일 수 있다. 마지막으로, 기록을 오픈 데이터(open data)로 바라보는 관점의 전환이 필요하다. 오픈 데이터의 핵심은 상호운용성으로, 서로 다른 데이터의 공개가 데이터 간의 상호 연결을 가능하게 한다(Open Knowledge Foundation, 2021). 기록의 궁극적 목적은 활용에 있고, 기록을 개방할수록 기록의 활용과 가치는 극대화된다. 기록의 개방은 국내외 연구자들 사이의 정보 공유를 활발하게 하고, 장기적으로 기록을 보존하고 공유하는 연구 협력체계를 만드는데 중요한 요소가 될 수 있다. 특히 일본군 '위안부' 기록은 파편화되어 존재하고, 기록의 발굴과 관리가 어려운 상황이기 때문에, 기록을 개방하고 공유함으로써 기록 보존의 효과적인 방법을 모색하고, 국내외 연구 협력을 이끄는 것이 필요하다.

참고문헌

국가기록원 (2013). 일본군 '위안부' 관련 자료.
출처: https://theme.archives.go.kr/next/nationalArchives/topicArchivesList.do?page=1&groupName=comport

국가기록원 (2014). 국가기록원고시 제2014-8호(시행 2014. 12. 26.).
출처: https://law.go.kr/LSW/admRulInfoP.do?admRulSeq=2200000025361

국가기록원 (2014). 국가지정기록물이란. 출처: https://theme.archives.go.kr/next/nationalArchives/archiveIntro.do

권미현 (2007). 강제동원 구술자료의 관리와 활용: 일제강점하강제동원피해진상규명위원회 소장 구술자료를 중심으로. 기록학연구, 16, 305-341.

김수현, 이성숙 (2020). RiC-CM을 적용한 영구기록물 기술방안 연구. 한국기록관리학회지, 20(1), 115-137. https://doi.org/10.14404/JKSARM.2020.20.1.115

김정현 (2020). 한중일의 일본군 '위안부' 기록물 발굴성과와 과제: 역사수정주의와 보편적 인권의 길항. 한일관계사연구, 69, 185-224. https://doi.org/10.18496/kjhr.2020.08.69.185

김학래 (2017). 지식그래프, 서울: 커뮤니케이션북스.

김학래 (2021). FAIR 원칙: 데이터 관점의 디지털 아카이브 구현을 위한 고려사항, 한국기록관리학회지, 20(1), 159-175. https://doi.org/10.14404/JKSARM.2021.21.2.155

남영주 (2017). 일본군 '위안부' 기록물 관리기관의 기억재현과 기억의 확장: 민족과 여성 역사관의 사례를 중심으로. 인문사회 21, 8(3), 129-148. https://doi.org/10.22143/HSS21.8.3.8

박선희 (2019). 기록물 맥락정보 향상 및 통합시스템 개발에 관한 연구: RiC-CM 및 RiC-O를 중심으로. 기록과 정보・문화 연구, 9, 55-96.

박지영 (2017). ISAD(G)에서 RiC-CM으로의 전환에 관한 연구. 한국기록관리학회지, 17(1), 93-115, https://doi.org/10.14404/JKSARM.2017.17.1.093

봉지현, 남영준 (2019). 일본군 '위안부' 구술기록의 관리를 위한 메타데이터 요소 선정에 관한 연구. 한국기록관리학회지, 19(1), 225-250. https://doi.org/10.14404/JKSARM.2019.19.1.225

서연수, 남연화, 박지원, 엄소영, 김용 (2016). 일본군 '위안부' 관련 기록물의 통합관리를 위한 메타데이터 스키마 개발에 관한연구, 한국기록관리학회지, 16(3), 99-129. https://doi.org/10.14404/JKSARM.2016.16.3.099

서울기록원 (2019). 2016~2018년 서울대 정진성 연구팀이 서울시의 지원을 받아 수집한 일본군 '위안부' 기록. 출처: https://archives.seoul.go.kr/contents/comfort-women

서울대 인권센터 정진성 연구팀 (2018). 끌려가다, 버려지다, 우리 앞에 서다 1. 서울: 푸른역사

서울대 인권센터 정진성 연구팀 (2018). 끌려가다, 버려지다, 우리 앞에 서다 2. 서울: 푸른역사

서현주 (2016). 2006~2016년간 일본군위안부 연구의 성과와 전망. 동북아역사논총, (53), 197-222.

성평등 아카이브 (2019). 일본군 위안부 관련 자료. 출처: http://www.genderarchive.or.kr/multi-collections/multi-collections/show/id/20

신미라, 김익한 (2019). RiC을 적용한 아카이브 시스템 데이터 모델링 연구. 한국기록관리학회지, 19(1), 23-67. https://doi.org/10.14404/JKSARM.2019.19.1.023

신혜수 (2021). 일본군 '위안부' 기록물의 세계기록유산 등재를 위해 ❷ 피해자의 아픔과 시민들의 노력 담긴 2744점의 역사. 출처: https://www.unesco.or.kr/data/unesco_news/view/780/1273/page/0?

윤지현 (2020). 아카이브 중심의 전쟁과여성인권박물관. 한국기록관리학회지, 20(4), 237-243. https://doi.org/10.14404/JKSARM.2020.20.4.237

이나영 (2010). 일본군 위안부 운동 - 포스트/식민국가의 역사적 현재성. 아세아연구, 53(3), 41-78.

이유경, 김학래 (2020). 1997 외환위기 지식그래프: 디지털 아카이브의 관계 중심적 접근. 한국기록관리학회지, 20(4), 1-17. https://doi.org/10.14404/JKSARM.2020.20.4.001

일본군 '위안부' 문제 연구소 (2020). 아카이브814. Available: https://www.archive814.or.kr/

전예지, 이혜원 (2020). RiC-CM v0.2 분석을 통한 온톨로지 모델링에 관한 연구. 한국기록관리학회지, 20(1), 139-158. https://doi.org/10.14404/JKSARM.2020.20.1.139

전쟁과 여성 인권 박물관 (2020). 수요시위 아카이브 컬렉션. 출처: https://www.archivecenter.net/wednesdaydemo/archive/Collection.do

전쟁과 여성 인권 박물관 (2020). 수요시위 아카이브. Available: https://www.archivecenter.net/wednesdaydemo

정의기억연대 (2018). 일본군 성노예제란? 출처: https://womenandwar.net/kr/what-is/

정회명, 이성숙 (2021). 디지털 환경에서 기록물 맥락 기술을 위한 Records in Contexts-Ontology(RiC-O) 적용 연구. 한국기록관리학회지, 21(2), 23-48. https://doi.org/10.14404/JKSARM.2021.21.2.023

한국기록전문가협회 (2020). [성명서] 나눔의 집의 '위안부' 기록 방치・절멸 행위에 대한 기록전문가 단체의 입장. 출처: https://www.archivists.or.kr/1636

Archives Nationales France (2021). Archives Nationales. Available: https://www.archives-nationales.culture.gouv.fr/en/web/guest/home

Borst, W. N. (1997). Construction of Engineering Ontologies for Knowledge Sharing and Reuse. Enschede: Centre for Telematics and Information Technology (CTIT).

Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, L., Umbrich, J., & Wahler, A. (2020). Introduction: what is a knowledge graph?. In Knowledge Graphs, New York City: Springer, 1-10.

Freire, N., Charles, V., & Isaac, A. (2018). Evaluation of Schema.org for Aggregation of Cultural Heritage. Preceedings of 15th International Conference on Extended Semantic Web Conference, Heraklion, Greece.

Freire, N., Robson, G., Howard, J. B., Manguinhas, H., & Isaac, A. (2020). Cultural heritage metadata aggregation using web technologies: IIIF, Sitemaps and Schema.org. International Journal on Digital Libraries, 21(1), 19–30. https://doi.org/10.1007/s00799-018-0259-5

Gruber, T. R. (1994). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 199-220. https://doi.org/10.1006/knac.1993.1008

Guha, R.V., Brickly, D., & Macbeth, S. (2016). Schema.org: Evolution of Structured Data on the Web. Communications of the acm, 59(2), 44-51. https://doi.org/10.1145/2844544

Han, M. K., Cole, T. W., Lampron, P., & Sarol, M. J. (2015). Exposing Library Holdings Metadata in RDF Using Schema.org Semantics. Proceedings of International Conference on Dublin Core and Metadata Applications, São Paulo, Brazil.

ICA EGAD (2016). Records in Contexts A Conceptual Model for Archival Description draft v0.1. Available: https://www.ica.org/sites/default/files/RiC-CM-0.1.pdf

ICA EGAD (2019). Records in Contexts − Ontology. Available: https://www.ica.org/en/records-in-contexts-ontology

ICA EGAD (2021). RiC-O projects and tools. Available: https://ica-egad.github.io/RiC-O/projects-and-tools.html

Jett, J., Cole, T., W., Han, M. K., & Szylowicz, C. (2017). Linked Open Data (LOD) for Library Special Collections. Proceedings of JCDL ’17 The 17th ACM/IEEE-CS Joint Conference on Digital Libraries, Toronto, Canada.

Lampron, P., Mixter, J., & Han M. K. (2016). Challenges of Mapping Digital Collections Metadata to Schema.org: Working with CONTENTdm. In Metadata and Semantics Research. New York City: Springer, 181-186.

Matienzo, M. A., Roke, E. R., & Carlson, S. (2017). Creating a Linked Data-Friendly Metadata Application Profile for Archival Description. Proc. Int’l Conf. on Dublin Core and Metadata Applications 2017, 112-116.

Memobase (2001). Memoriav Memobase. Available: https://memobase.ch/fr/start

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11-33. https://doi.org/10.1109/JPROC.2015.2483592

Open Knowledge Foundation (2021). Open Data Handbook. What is Open Data? Available: https://opendatahandbook.org/guide/en/what-is-open-data/

Schema.org (2021). Organization of Schemas. Available: https://schema.org/docs/schemas.html

Shin, H. (2021). Voices of the “Comfort Women”: The Power Politics Surrounding the UNESCO Documentary Heritage. The Asia-Pacific Journal, 19(5), 1-19.

Social Networks and Archival Context Cooperative (2021). Social Networks and Archival Context. Available: https://snaccooperative.org/

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge Engineering: Principles and methods. Data & Knowledge Engineering 25, 161-197.

UNESCO (2021). Memory of the World. Available: https://en.unesco.org/programme/mow

Vrandečić, D. (2012). Wikidata: A new platform for collaborative data collection. In Proceedings of the 21st international conference on world wide web, 1063-1064, https://doi.org/10.1145/2187980.2188242

W3C Schema Architypes Community Group (2015). W3C Schema Architypes Community Group. Available: https://www.w3.org/community/architypes/

W3C Schema Bib Extend Community Group (2015). W3C Schema Bib Extend Community Group. Available: https://www.w3.org/community/schemabibex/

Zou, X. (2020). A survey on application of knowledge graph. In Journal of Physics. Paper presented at 4th International Conference on Control Engineering and Artificial Intelligence (CCEAI 2020), Singapore.

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

Bong, Ji Hyeon & Nam, Young Joon (2019). A Study on the Design of Metadata Elements for Management of Oral History Archives about Sexual Slavery by Japan's Military. Journal of Korean Society of Archives and Records Management, 19(1), 225-250. http://doi.org/10.14404/JKSARM.2019.19.1.225

Chung, Chin-Sung Team at SNU Human Rights Center (2018). To be taken, to be abandoned, to stand before us 1. Seoul: Purunoksa.

Chung, Chin-Sung Team at SNU Human Rights Center (2018). To be taken, to be abandoned, to stand before us 2. Seoul: Purunoksa.

Gender Archive (2019). Japanese 'Comfort Women' Collection. Available: http://www.genderarchive.or.kr/multi-collections/multi-collections/show/id/20

Jeon, Ye Ji & Lee, Hyewon (2020). A Study on the Ontology Modeling by Analyzing RiC-CM v0.2. Journal of Korean Society of Archives and Records Management, 20(1), 139-158. https://doi.org/10.14404/JKSARM.2020.20.1.139

Jeong, Hoemyeong & Lee, Sungsook (2021). A Study on the Application of Records in Contexts-Ontology(RiC-O) for the Description of Archives Contexts in a Digital Environment. Journal of Korean Society of Archives and Records Management, 21(2), 23-48, https://doi.org/10.14404/JKSARM.2021.21.2.023

Kim, Haklae (2021). FAIR Principles: Considerations for Implementing Digital Archives from a Data Perspective, Journal of Korean Society of Archives and Records Management, 21(2), 155-172. https://doi.org/10.14404/JKSARM.2021.21.2.155

Kim, Jeonghyun (2020). Achievements and Tasks of the excavation of 'Japanese military sexual slavery' records in Japan-China-Korea. The Korea-Japan Historical Society, 69, 185-224. https://doi.org/10.18496/kjhr.2020.08.69.185

Kim, Soohyun & Lee, Sungsook (2020). A Study on Archive Description Using RiC-CM. Journal of Korean Society of Archives and Records Management, 20(1), 115-137. https://doi.org/10.14404/JKSARM.2020.20.1.115

Korea Association of Archivists (2020). [Statement] Archivists' group position for neglection and annihiliation of House of Sharing's Japanese 'Comfort Women' records. Available: https://www.archivists.or.kr/1636

Korean Council for Justice and Remembrance for the Issues of Military Sexual Slavery by Japan(Korean Council). (2018), What is Japanese Military Sexual Slavery System? Retrieved May 8th, 2021, from https://womenandwar.net/kr/what-is-japanese-military-sexual-slavery-system/

Kwon, Mi-Hyun (2007). Management and Use of Oral History Archives on Forced Mobilization −Centering on oral history archives collected by the Truth Commission on Forced Mobilization under the Japanese Imperialism Republic of Korea-. Journal of Korean Society of Archives and Records Management, 16, 305-341.

Lee, Na-Young (2010). Womens Movement for/on Comfort Women: Historical Present in the Context of Postcolonial Nation-State. The Journal of Asiatic Studies, 53(3), 41-78.

Lee, Yu-kyeong & Kim, Haklae (2020). A Knowledge Graph of the Korean Financial Crisis of 1997: A Relationship-Oriented Approach to Digital Archives. Journal of Korean Society of Archives and Records Management, 20(4), 1-17, https://doi.org/10.14404/JKSARM.2020.20.4.001

Nam, Young Joo (2017). Memory Replay and Memory Expansion of the Japanese Military Sexual Slavery Records Management Institutions: On the Cases of National Women's Historical Hall. The Journal of Humanities and Social science (HSS21), 8(3), 129-148. https://doi.org/10.22143/HSS21.8.3.8

National Archives of Korea (2013). Japanese 'Comfort Women' records. Available: https://theme.archives.go.kr/next/nationalArchives/topicArchivesList.do?page=1&groupName=comport

National Archives of Korea (2014). National Archives Notice No. 2014-8(Implementation 2014. 12. 26.). Available:

https://law.go.kr/LSW/admRulInfoP.do?admRulSeq=2200000025361

National Archives of Korea (2014). What is Nation-designated Archives? Available: https://theme.archives.go.kr/next/nationalArchives/archiveIntro.do

Park, Sun-hee (2019). A Study on Improving Record Contextual Information and Developing Integrated System: Focusing on RiC-CM and RiC-O. The Korean Journal of Archival, Information and Cultural Studies, 9, 55-96.

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. https://doi.org/10.14404/JKSARM.2017.17.1.093

Research Institute on Japanese Military Sexual Slavery (2020). Archive814. Available: https://www.archive814.or.kr/

Seo, Hyunju (2016). 2006-2016 Research Progress on the Japanese Military Comfort Women Issue and a Future Outlook: Focusing on Historical Studies in South Korea. Dongbuga Yeoksa Nonchong, (53), 197-222.

Seo, Yeon-Su, Nam, Yeon-Hwa, Park, Ji-Won, Um, So-Young, & Kim, Yong (2016). A Study on the Development of a Metadata Schema for the Records and Archives on the Military Sexual Slavery by Japan. Journal of Korean Society of Archives and Records Management, 16(3), 99-129. https://doi.org/10.14404/JKSARM.2016.16.3.099

Seoul Metropolitan Archives (2019). Japanese 'Comfort Women' records collected by Chung Jin-sung, a research team of Seoul National University, from 2016 to 2018 with the support of the Seoul Metropolitan Government. Available: https://archives.seoul.go.kr/contents/comfort-women

Shin, Heisoo (2021). To register for World Heritages of Japanese 'Comfort Women' archives ❷ 2,744 pieces of history containing the pain of the victim and the efforts of the citizens. Retrieved July 12, 2021, Available: https://www.unesco.or.kr/data/unesco_news/view/780/1273/page/0?

Shin, Mira & Kim, Ikhan (2019). A Study in the Data Modeling for Archive System Applying RiC. Journal of Korean Society of Archives and Records Management, 19(1), 23-67, https://doi.org/10.14404/JKSARM.2019.19.1.023

Women and War Museum (2020). Wednesday Demo Archive Collection. Available: https://www.archivecenter.net/wednesdaydemo/archive/Collection.do

Women and War Museum (2020). Wednesday Demo Archive. Available: https://www.archivecenter.net/wednesdaydemo

Youn, Jihyun (2020). War and Women's Human Rights Museum: Archives are Key. Journal of Korean Society of Archives and Records Management, 20(4), 237-243. https://doi.org/10.14404/JKSARM.2016.16.3.09910.14404/JKSARM.2020.20.4.237